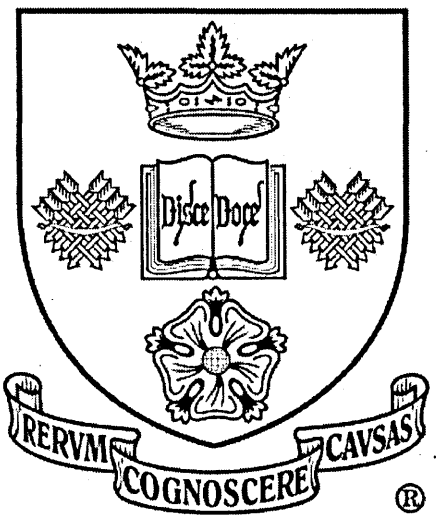
Activity fingerprints in DNA based on a structural analysis of sequence information

A study submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy



at
The University of Sheffield

by
Linda Hirons

Department of Chemistry and
Department of Information Studies

October 2005

Acknowledgements

Thank you to Professor Chris Hunter, my supervisor from the Department of Chemistry, and Professor Peter Willett, my supervisor from the Department of Information Studies. Also, thanks to Dr. Eleanor Gardiner for her constant support throughout this research and her many inspiring ideas.

I acknowledge Dr. Sean Eddy, since without his well constructed HMMER package this research would have been much more challenging. My appreciation goes to both Dr Anders Pedersen and Dr Dave Ussery for their datasets and kind email correspondence.

Finally, cheers to my friends and family for all their positive encouragement and to my niece and nephews who will always keep me smiling.

This research was funded by BBSRC.

Abstract

The function of a DNA sequence is commonly predicted by measuring its nucleotide similarity to known functional sets. However, the use of structural properties to identify patterns within families is justified by the discovery that many very different sequences have similar structural properties. The aim of this thesis is to develop tools that detect any unusual structural characteristics of a particular sequence or that identify DNA structure-activity fingerprints common to a set.

This work uses the Octamer Database to describe DNA. The database's contents are split into two categories: those parameters that describe minimum energy structure and those that measure flexibility. Information from both of these categories has been combined to describe structural tendencies, offering an alternative measure of sequence similarity.

A structural DNA profile gives a graphical illustration of how a parameter from the Octamer Database varies across either a single sequence's length or across a set of sequences. Profile Manager is an application that has been developed to automate single sequence profile generation and is used to study the A-tract phenomenon. The use of profiles to explore patterns in flexibility across a set of pre-aligned promoters is then investigated with interesting transitions in decreasing twist flexibility discovered.

Multiple sequence queries are harder to solve than those of single sequences, due to the inherent need for the sequences to be aligned. It is only under rare circumstances that sequences are pre-aligned by an experimentally determined position. More commonly a multiple alignment must be generated. An extended, structure-based, hidden Markov model technique that successfully generates structural alignments is presented. Its application is tested on four DNA protein binding site datasets with comparisons made to the traditional sequence method. Structural alignments of two out of the four datasets were comparable in performance to sequence with useful insights into underlying structural mechanisms.

Contents

Chapter 1:
Introduction

Molecules of Deoxyribonucleic Acid (DNA) store most of the hereditary information belonging to a particular living organism. Recent studies have suggested that molecules other than DNA may also transfer hereditary information, as proteins such as prions have recently been identified as genetic elements (Bussard, 2005). The majority of DNA is located in the nucleus of every cell, with the exception of red blood cells, in our body. A small proportion of DNA is found within an organelle called the mitochondria. This mitochondrial DNA differs from the nuclear DNA as it is only inherited maternally (Chen and Butow, 2005).

DNA defines who we are by encoding the structures of proteins and enzymes that our body manufactures. The production of proteins is very important, controlling the state of a cell and processes such as muscle building, the digestion of food and synthesis of hormones. The segments of DNA that encode protein structure are known as our genes and make us unique individuals that carry a unique combination of our parents' genetic material. Recently, the concept of genetic imprinting has been recognised, whereby the activation of a gene can be switched on or off depending on whether it has been inherited maternally or paternally (Mager and Bartolomei, 2005). Imprinting is one example of epigenetics, which is how the function of a gene can be altered without any changes being made to the DNA sequence. Understanding how DNA works is very important in identifying many genetic diseases and enables us to have a greater knowledge of what controls who we are and the environment around us.

It would be a mistake to believe that the main bulk of our DNA is made up of genes. 95% of our DNA is non-coding and is often referred to as 'junk', since we do not understand its function. Discovering the functional purposes of some of this so-called 'junk DNA' has been likened to searching through "Heirlooms in the attic" and finding hidden gems (Johnston and Stormo, 2003). The development of computational tools to detect any unusual characteristics of a particular sequence or to identify patterns common to a set of sequences will therefore be valuable.

The use of structural properties to identify patterns within families of sequences is justified by the discovery that many very different sequences have similar structural properties (Gardiner et al., 2004). This means that by looking at the information hidden within the structure, similarities between DNA sequences will be found that would otherwise be unrecognised. Observing how the structure of sequences vary across their length will not only help predict unknown functions, but will also be a key to understanding the structural mechanisms involved in known functions (something that cannot be done by looking at a string of nucleotide letters). The aim of this work is to develop and use tools that analyse how the structure of DNA varies with its function, in order to identify structural patterns, known as activity fingerprints.

Chapter 2 describes the double helical structure of DNA and presents a well-known system of nomenclature used to describe subtle differences between the helical geometry of sequences. This research uses the Octamer Database (Gardiner et al., 2003) to encode the sequence-dependent structure of DNA. An octamer is a DNA sequence of nucleotide length eight ($x_1x_2x_3x_4x_5x_6x_7x_8$, where x equals A, C, G or T). The database contains structural properties describing the minimum energy conformation and flexibility of all unique octamers. Chapter 3 gives a detailed description of the Octamer Database and explores correlations between its parameters. In brief, the minimum energy structure is described by the base-step parameters and three ground state properties: energy, groove and RMSD. The flexibility is described by the force constants and partition coefficients. Chapter 4 presents a novel extension to the Octamer Database, combining the minimum energy of an octamer with its flexibility in order to calculate structural probabilities. This offers a novel way of comparing two sequences, allowing the dynamical structure of DNA to be studied.

Chapters 5 and 6 present the structural DNA profiles, which can be used to visualise activity fingerprints. A profile is a graphical illustration of how a structural parameter from the Octamer Database varies across either a single sequence's length or across a set of sequences. Chapter 5 looks at single sequence queries with presentation of an application to automate the generation of single sequence profiles (Profile Manager). Development of the graphical user interface is explained and program functionality is described with examples. Chapter 6 then explores multiple sequence queries and how they can be answered by multiple sequence profiles.

Application of the multiple sequence profiles is restricted to pre-aligned sequence datasets. It will not be possible to find structural patterns without an alignment method. A structural alignment tool is therefore needed. A method commonly used to generate sequence alignments is Hidden Markov Models (HMMs). Chapter 7 examines the HMM technique with some simple examples. The traditional model architecture used to analyse biological sequences is presented. Details are given on model construction, including explanation of the commonly used Viterbi algorithm, Forward algorithm and Baum-Welch procedure. Other topics covered are the alternative Simulated Annealing technique, model surgery, prior knowledge and sequence weighting. A survey of analogous structural HMM work is also performed.

Chapter 8 then presents a novel structural DNA alignment technique, which currently aligns sequences by a single minimum energy parameter (3-step roll). Flexibility is encoded within a model's prior knowledge, therefore considering the dynamical nature of DNA. Methods for assessing the performance and predictive ability of HMMs are presented and an artificial dataset applied in order to test the functionality of the technique before applying it to real data. Structural HMMs of four protein-DNA binding site datasets are then constructed, assessed and compared to their traditional sequence models in Chapter 9. Chapter 10 then summarises the conclusions made throughout this work, making suggestions for future research.

Chapter 2:
Deoxyribonucleic Acid

Over fifty years ago a breakthrough in scientific research occurred when the structure of DNA was discovered and published (Watson and Crick, 1953). Since then vast amounts of DNA research have been carried out with structural discoveries beyond the double helix continuously being made (Pearson, 2003), reflecting how much there is still to learn about this complicated molecule. By the 1970's, experimental techniques to determine the nucleotide sequence of a piece of DNA had been discovered. However it was not until about 1995 that technology had become advanced enough to deal with the size of the human genome (Olson, 1995). This resulted in an enormous scientific challenge, the human genome project (HGP). Finally on the 50th anniversary of Watson and Crick's discovery the HGP was completed (Collins et al., 2003a; Collins et al., 2003b; Frazier et al., 2003), providing a wealth of information to analyse. This has opened the doors to further understanding DNA structure and to discovering cures for numerous genetic diseases.

This chapter describes the double helical structure of DNA. The exact structure of a double helix is dependent upon its nucleotide sequence, therefore parameters that encode the geometry of a helix are needed. A standard system of nomenclature for such parameters has been agreed and is known as the Cambridge Accord (Diekmann, 1989). There are three categories of geometric descriptors: the base-pair parameters, base-step parameters and global parameters. The development of a computational model that can be used to accurately predict the geometry of any DNA double helix is reviewed.

2.1. The Double Helix

DNA is a biopolymer whose polymeric building blocks are the nucleotides. A nucleotide consists of a sugar, phosphate group and base (Figure 2.1). Note the labelling of carbon atoms around the sugar (Figure 2.1). The position at which the base is bound is numbered 1 (C1'). In DNA the sugar is 2'deoxyribose, ribose with oxygen

removed from position 2. The phosphate is connected to the C5' carbon of the sugar via a phosphate-ester bond. The sugars and phosphates of adjacent nucleotides join together to form a sugar-phosphate backbone, producing a single strand of DNA (Figure 2.2). The ends of a strand are labelled 3' and 5' based on the positioning of sugar atoms. This is known as the directionality of DNA.

Figure 2.1: *The Nucleotide. The polymeric building block of DNA, consisting of a phosphate group, sugar and base.*

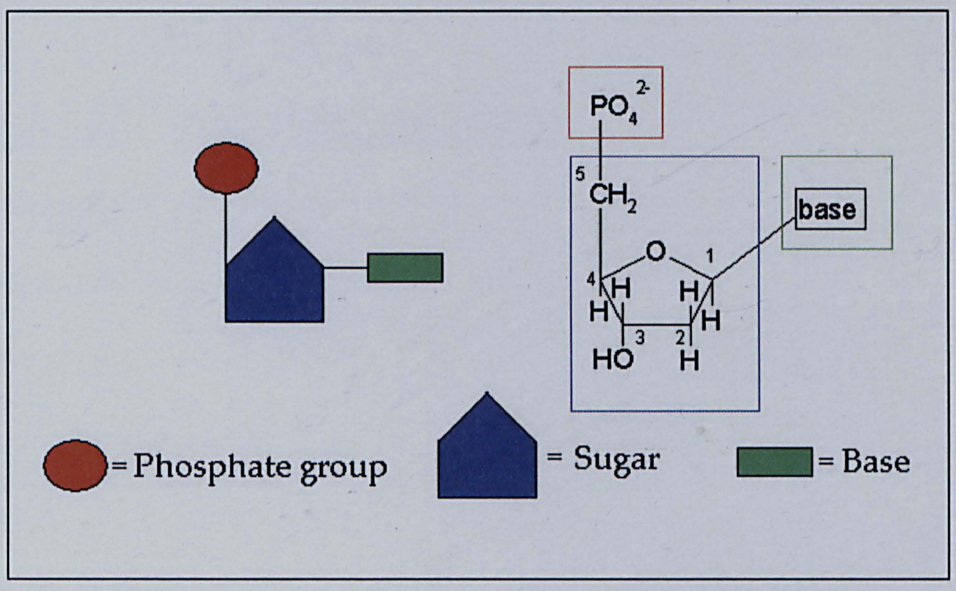
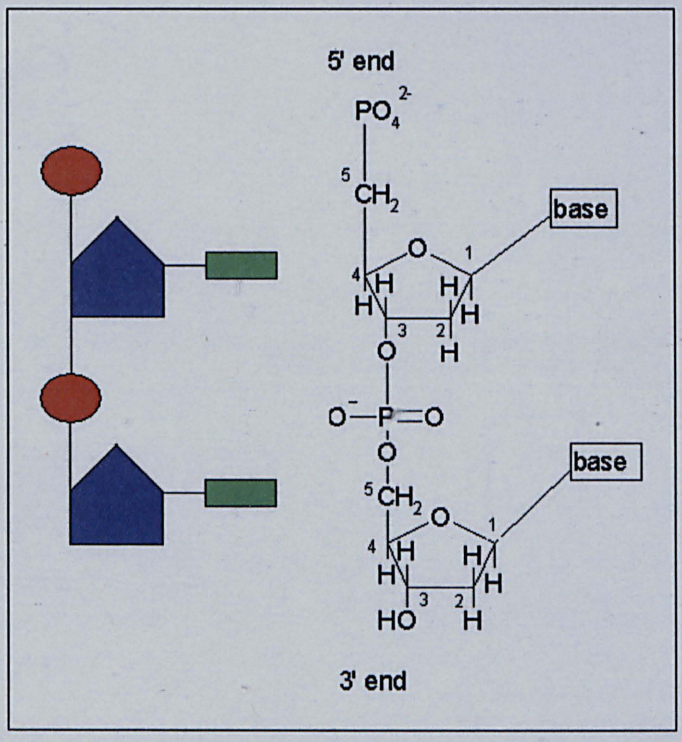
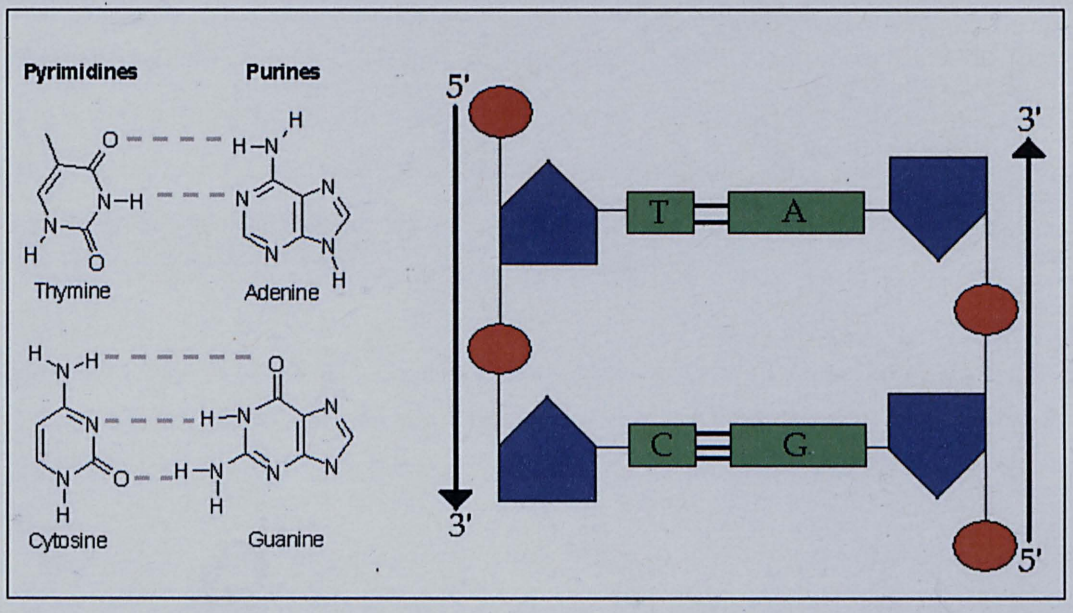


Figure 2.2: *The sugar-phosphate backbone. Nucleotides join together to form a single strand of DNA.*



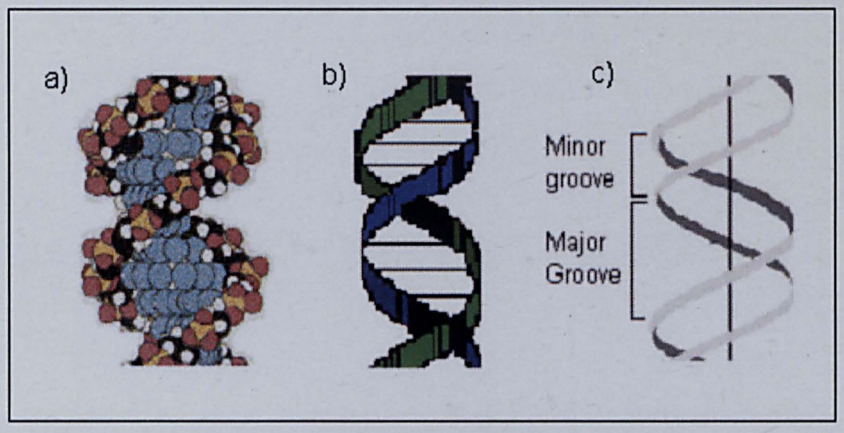
The heterocyclic amine bases are either pyrimidines (thymine, T and cytosine, C) or purines (adenine, A and guanine, G). The chemical structures of these four different bases are given in Figure 2.3. Hydrogen bonds are shown between purine-pyrimidine bases and form the Watson-Crick base pairs (A-T and G-C) (Watson and Crick, 1953). A-T and G-C are equal in length, enabling DNA to form a double stranded structure analogous to a ladder. The energy to break a G-C interaction is greater than that required to break an A-T one, due to three hydrogen bonds versus two. The 5' to 3' directionality of the two strands run in opposite directions, they are anti-parallel. The strict base pairing rules mean that the strands are complementary to one another. For this reason the sequences ATGCCA and TGGCAT are equivalent.

Figure 2.3: *The Watson-Crick base pairs and double stranded structure of DNA.*



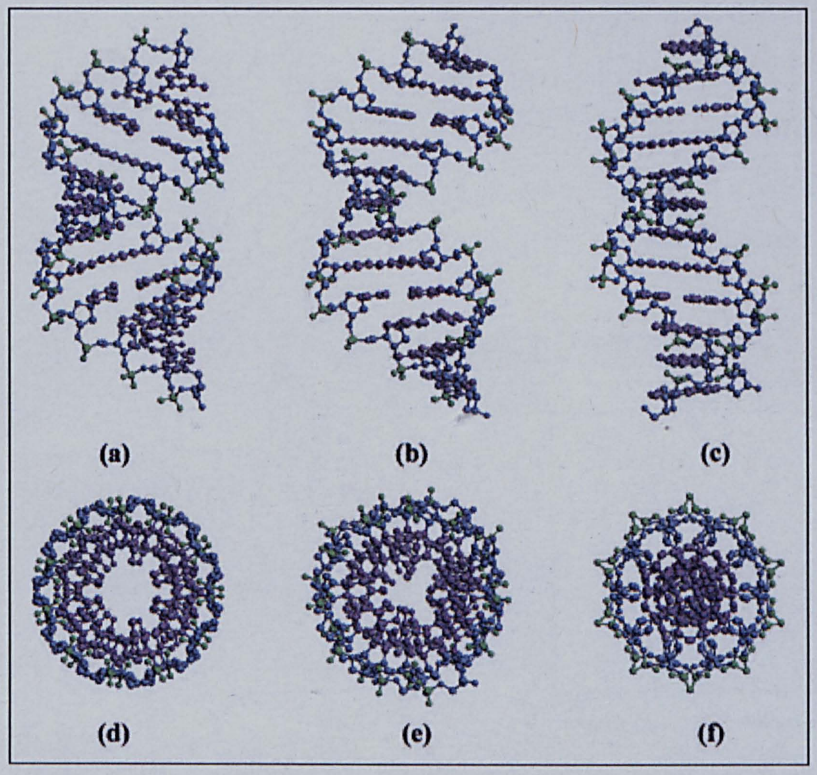
The double stranded structure is twisted to form a double helix (Watson and Crick, 1953), see Figure 2.4. The sugars and phosphate groups are hydrophilic and therefore highly soluble in the aqueous cellular environment. The bases however are hydrophobic and place conformational constraints on DNA in vivo. The formation of a double helix stabilises the structure by keeping the hydrophobic portion of the molecule in its interior, where solvent accessibility is kept to a minimum. There are two helical grooves (Figure 2.4c) that expose parts of the base pairs to the surrounding environment and that enable drugs and proteins to recognise and bind to specific sequences. The deeper groove is called the major groove and the smaller of the two is the minor groove.

Figure 2.4: *The double helix of DNA. (a) All atom representation. (b) Cartoon representation analogous to a twisted ladder. (c) The two grooves that expose parts of the base pairs.*



DNA tends to have 10 to 12 phosphates per helical turn and can be broadly split into three types: A, B and Z. A, B and Z have 11, 10 and 12 phosphates per turn respectively with A (Figure 2.5a) and B (Figure 2.5c) having right-handed helices and Z having a left-handed helix. In reality DNA takes on intermediate structures between these extreme cases, its exact three-dimensional structure being sequence dependent. For example the crystal structure of CATGGGCCCATG (Figure 2.5b) is an intermediate between idealised A-DNA and B-DNA.

Figure 2.5: *The sequence dependent structure of DNA (Ng et al., 2000). Crystal structure of A-DNA (a) viewed from the side and (d) from the top. Crystal structure of CATGGGCCCATG (b) viewed from the side and (e) from the top. Crystal structure of B-DNA (c) viewed from the side and (f) from the top.*



2.2. The Cambridge Accord

Parameters describing the geometry of DNA helices have been devised, in order to investigate how the shape of a double helix varies with its base-pair sequence. In 1988 a meeting held in Cambridge (Dickerson et al., 1988) set down an agreement across the scientific community on the nomenclature of the translations and rotations required to successfully describe a helix. This agreement is known as the Cambridge Accord (Diekmann, 1989). There are three classes of motions within a helix: base-pair motions, base-step motions and global motions (see Table 2.1). Each motion is described by six degrees of freedom (three translations and three rotations).

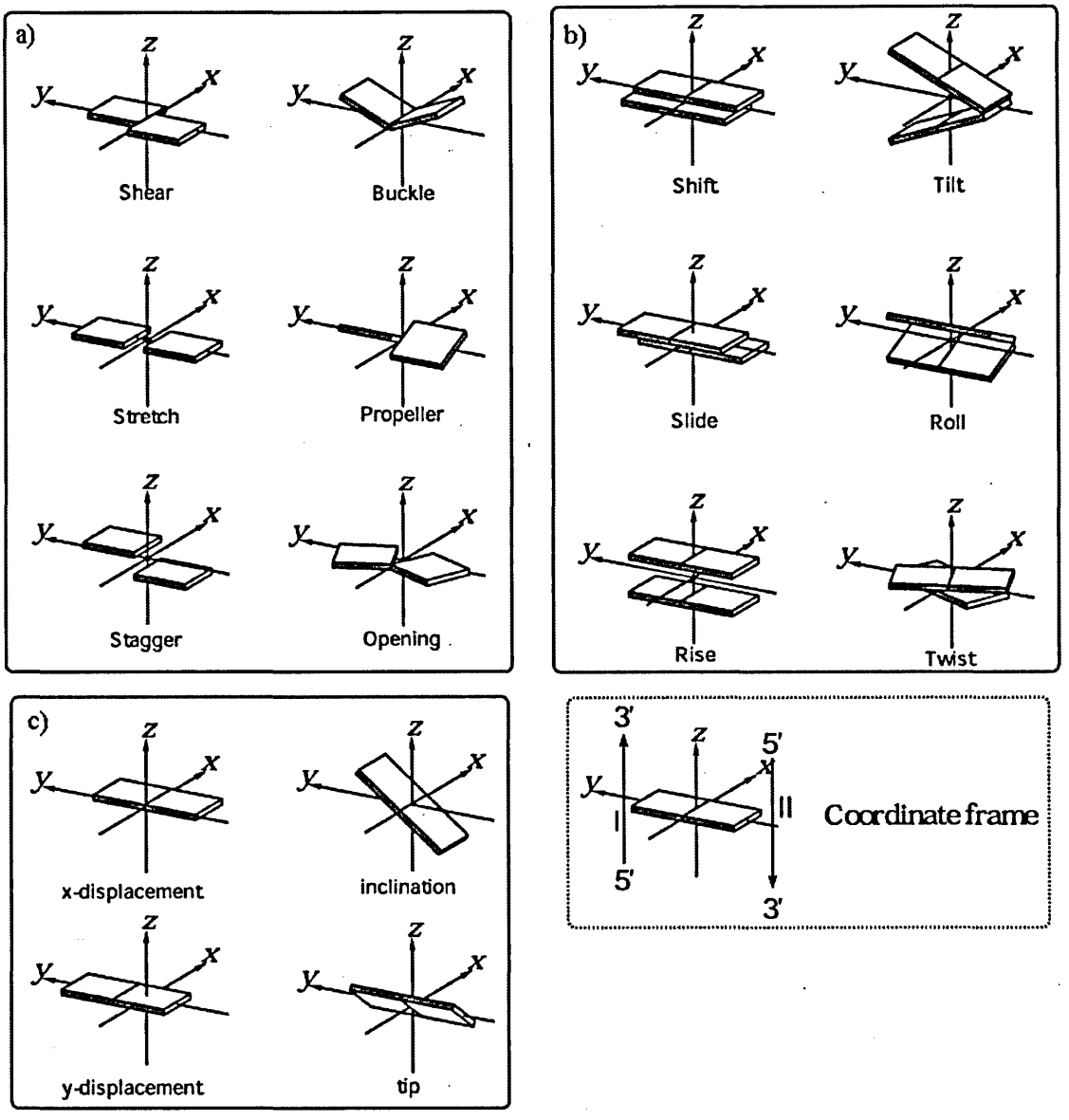
Table 2.1: *The Cambridge Accord Nomenclature (Diekmann, 1989).*

Motion	Translation Axis			Rotation Axis		
	x	y	z	x	y	z
Pair	Shear	Stretch	Stagger	Buckle	Propeller	Opening
Step	Shift	Slide	Rise	Tilt	Roll	Twist
Global	x-displacement	y-displacement	$Rise_g$	Inclination	Tip	$Twist_g$

The base-pair parameters describe the geometry of the two bases within a single base-pair, see Figure 2.6a (Lu & Olson, 2003). The three rotations are buckle, propeller and opening. The three translations are shear, stretch and stagger. Only two of the base-pair parameters have been found to vary significantly: the x-rotation buckle and y-rotation propeller (Yanagi et al., 1991). The base-step parameters (Figure 2.6b) describe the geometry of two adjacent base-pairs. The three rotations are twist, roll and tilt. The three translations are rise, slide and shift. Finally, the global parameters (Figure 2.6c) describe the geometry of the base-pairs relative to a global reference frame.

The positive directions along the x, y and z-axes are shown in the coordinate frame of Figure 2.6. The x-direction is along the short axis of the base-pair, the y-direction is along the long axis of the base-pair and the z-direction is perpendicular to the base-pair. The Cambridge Accord adopts a right-hand rule for the direction of the rotations, meaning that clockwise rotations about an axis are positive and anti-clockwise rotations are negative.

Figure 2.6: *(a) The base-pair, (b) base-step and (c) global parameters (Lu and Olson, 2003)*



The Cambridge University Engineering Department Helix Computation Scheme (CEHS) (El-Hassan and Calladine, 1995) calculates the base-pair and base-step parameters in agreement with the Cambridge Accord. Consider the calculation of the base-step parameters. Firstly, the location of the two adjacent base-pairs forming the base-step must be described. The positioning in space of a single base-pair is encoded by its own individual reference frame, known as a base-pair triad. A base-pair triad provides a set of three axes (x, y and z) with a specified origin from which the orientation of the base-pair can be deciphered. The difference in geometry between two

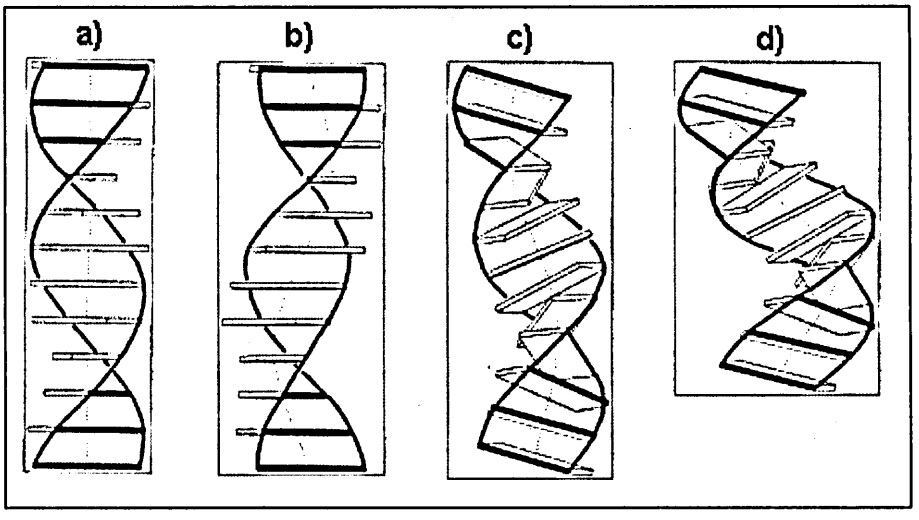
adjacent base-pairs could therefore be considered as the rotations and translations required to transform the triad of the first base-pair into that of the second base-pair. However, a problem occurs with this approach, as the measurements observed will depend on the observer's frame of reference. In other words, the resulting base-step parameters will differ with which base-pair is chosen as the reference. For this reason the CEHS introduced the concept of mid-step triads. A mid-step triad, as its name suggests, is a reference frame located between the two base-pairs in such a way that the base-step parameters are identical despite the direction in which the step is read. For full mathematical details of the base-pair triads and how the associated mid-step triad is calculated see the CEHS (El-Hassan and Calladine, 1995).

The CEHS scheme was used to analyse a database containing the X-ray crystal structures of 60 DNA oligomers (El-Hassan and Calladine, 1996). The variation in slide for each dinucleotide step was assessed via a frequency plot with the standard deviation being taken as a flexibility measure (slide mobility). Some steps were found to be rigid with a tendency for a single value of slide (e.g., AA), some were flexible with a wider slide range (e.g., CA) and others were bistable with a bimodal slide frequency distribution (e.g., GG). The mean propeller of a step was identified as being inversely proportional to its slide mobility, due to high propeller acting as a "steric interlock" that causes low slide mobility (El-Hassan and Calladine, 1996).

The Structure and Conformation of Helical Nucleic Acids Analysis program, SCHNAaP (Lu et al., 1997a) implements and extends the CEHS scheme. Along a sequence it generates the 18 parameters presented in Table 2.1 from the atomic coordinates. Backbone descriptors are also calculated, which include a variety of torsion angles and groove widths. Mismatched base-pairs and subtle variations of the base structures, such as absent methyl groups, can be dealt with. The output includes base stacking illustrations and polymorphic family assignment. A program that carries out the reverse procedure, rebuilding structure from base-pair and step parameters, exists: SCHNArP (Lu et al., 1997b). SCHNArP offers a valuable way of comparing and evaluating structures predicted by different models. SCHNAaP and SCHNArP, collectively referred to as SCHNAP, have now been replaced and superseded by 3DNA (Lu and Olson, 2003).

Base-stacking diagrams are useful when analysing the transitions required to convert the crystal structure of B-DNA (refer back to Figure 2.5c on page 7) into A-DNA (refer back to Figure 2.5a on page 7). A two-stage conversion involving a change in slide and a change in roll is presented in Figure 2.7 (Dickerson and Ng, 2001). The effect of negative slide upon a helix applied evenly at each base-step can be seen from the difference between Figure 2.7a and b. Likewise the effect of positive roll can be seen from the differences between Figure 2.7a and c. Applying both of these changes simultaneously results in transforming idealised B-DNA of no slide or roll into idealised A-DNA having a slide of –1.5 Angstroms and roll of 12°

Figure 2.7: *Converting B-DNA to A-DNA by uniform changes in base-step slide and base-step roll (Dickerson and Ng, 2001). (a) Idealised B-DNA with no slide or roll. (b) Intermediate with uniform slide of –1.5 Angstroms. (c) Intermediate with uniform roll of 12°. (d) Idealised A-DNA with slide of –1.5 Angstroms and roll of 12°.*



2.3. Base-stacking model

The stacking interactions of the 16 possible base-steps were calculated as a function of slide, roll, twist and propeller (Hunter, 1993). Shift and tilt were set to zero and rise was altered to ensure van der Waals' contacts between the base-pairs. For simplicity the sugar phosphate backbone was ignored and the dielectric constant set to unity. Energy contour plots of slide versus roll were used to examine the conformational preferences. Complementary steps gave identical results, leading to ten unique steps with energy minimum structures that agreed qualitatively with experiment

(Hunter, 1993). The properties of certain base-pair steps were found to be related to the properties of the constituent base-pairs. For example, the incompatible conformations of TA and AT account for the strong preference of DNA to unwind when the TATA sequence is present. This supports the importance of the TATA-box as a core promoter element in transcription initiation.

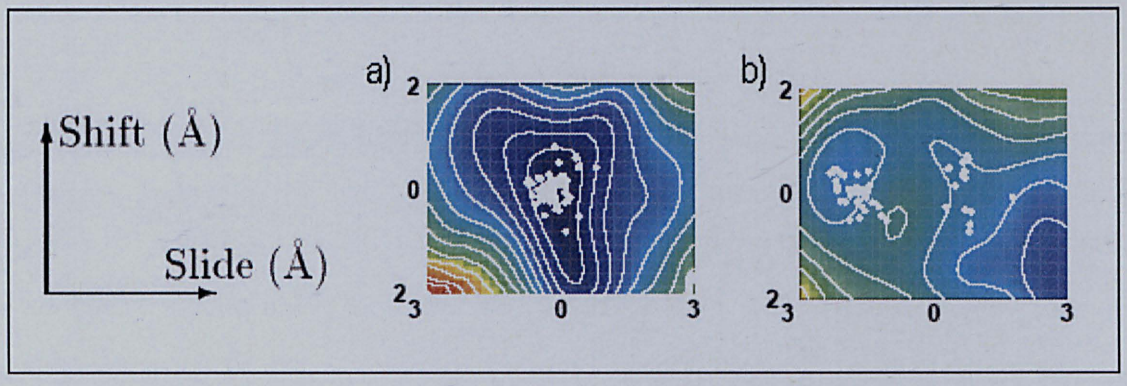
Correlations between the experimental base-step parameters of the oligomer database (El-Hassan and Calladine, 1996) and those predicted by Hunter's stacking interaction model have been analysed (Hunter and Lu, 1997). The dielectric constant was investigated with four being more accurate than unity. Calculated values of rise, roll and tilt correlated well with the experimental data, meaning these descriptors are determined solely by the base stacking interactions. Upon further inspection it was found that these three degrees of freedom were almost entirely explained by just the van der Waals' interactions, due to their association with the vertical separation between base-pairs. Slide and shift were not so well predicted, with electrostatic interactions being important in explaining the range of values observed. The precise values however are constrained by both the sugar phosphate backbone and the conformations of neighbouring steps, two factors that needed further investigation. Twist has a very poor correlation with the base stacking interactions and is thought to be entirely dependent upon the backbone.

A semi-flexible rod model of the backbone constraint has been constructed (Packer and Hunter, 1998). Two parameters are used, the mean backbone length for a step and the difference in backbone lengths. Both of these parameters can be accurately calculated from slide, shift and propeller. Twist can now be accurately modelled, confirming that it is totally dependent on the backbone. Slide and shift cannot be predicted, due to their context dependence upon neighbouring steps.

The 16 dinucleotide energy contour maps of slide versus shift with optimised values of twist, roll, tilt and rise have been investigated (Packer et al., 2000a). Certain steps were found to be bistable, meaning that they possess two or more distinct energy minima in their slide-shift conformational energy maps. Minima were defined as separate if their slide values differed by at least 1 Angstrom. The bistability of GG and CC can be seen clearly by the presence of two energy minima (see Figure 2.8b) in

comparison to the single global minima of the AA step (Figure 2.8a). However the known bistability of GC and CG is not observed, since it is a property of sequence context effects not described at the dinucleotide level. Slide flexibility was assessed by fitting quadratic equations to energy minima paths and the results agreed well with slide mobility (El-Hassan and Calladine, 1996).

Figure 2.8: *Energy contour maps of slide versus shift in Angstroms (Packer et al., 2000b) for the dinucleotide (a) AA that possesses a single distinct energy minimum and (b) GG which is clearly bistable.*



Contour plots analogous to those just described have also been constructed for all tetranucleotides (Packer et al., 2000b): these successfully describe the context dependent effects, providing accurate predictions of slide and shift. This work proposed a two-term model for calculating the energy of an oligomer of length N ($E^{N}_{oligomer}$), which is shown as the first two terms of Equation 2.1 (the base-step energies and step junction contributions). Some steps are context independent, due to neighbouring steps with compatible conformational properties and others have strong context dependence (Table 2.2). Neighbouring slide values along a sequence are strongly correlated. Neighbouring shift values along a sequence are anti-correlated. This means that slide has a tendency to be similar along a sequence and shift has a tendency to alternate.

Table 2.2: *Context Dependence of Dinucleotides as classified by Packer et al. (2000b).*

Classification	Steps
Context Independent	AA/TT, AT, TA
Weakly Context Dependent	AC/GT, AG/CT, CA/TG, GA/TC
Strongly Context Dependent	CG, GC, CC/GG

An extension to the above model was used to make predictions about 30 oligomers (Packer and Hunter, 2001). A third term was introduced to describe base-backbone interactions via a penalty function to account for steric clashes between a base and furanose sugar (E^n_{sugar}). See Equation 2.1, where E^n_{step} is the energy of the nth step and $E^n_{junction}$ is the step junction contribution.

$$E^N_{oligomer} = \sum_{n=1}^{N-1} E^n_{step} + \sum_{n=2}^{N-2} E^n_{junction} + \sum_{n=1}^{N} E^n_{sugar} \qquad \textbf{Equ. 2.1}$$

A genetic algorithm was used to search for an oligomer's global minimum energy structure, followed by a grid search to identify local minima. 24 of the 30 sequences had their structures accurately predicted, with three of the remaining having their differences accounted for by crystal packing in the solid state.

Chapter 3:
The Octamer Database

Conformational energy maps of central base-step slide and shift have been stored and used to calculate the structural properties of the 32,896 unique octamers, creating the Octamer Database (Gardiner et al., 2003). The database's contents can be split into two categories: the parameters that describe an octamer's minimum energy structure and those that describe an octamer's flexibility.

3.1. Describing the minimum energy structures

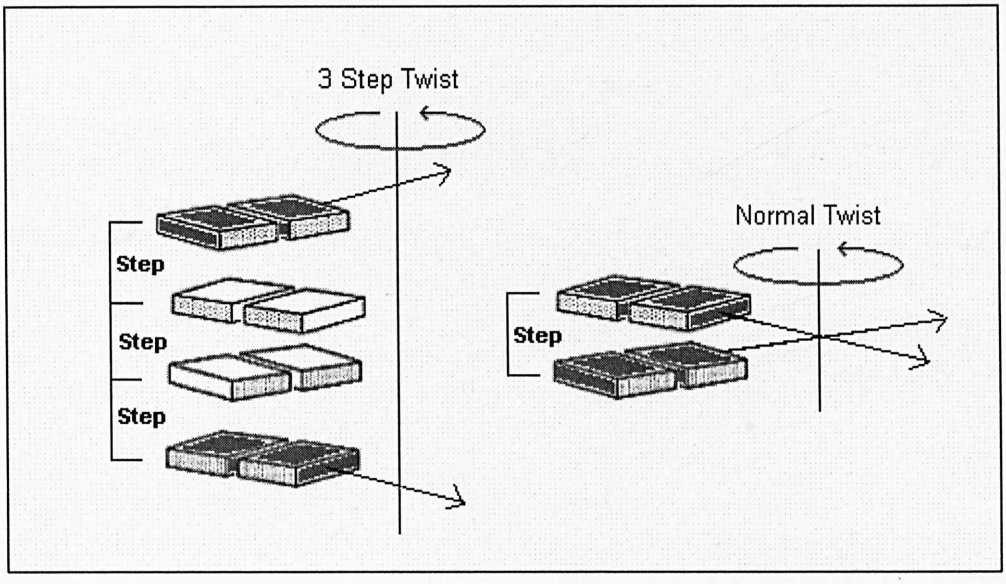
Minimum energy structures are described by the parameters in Table 3.1 (the six minimum energy step parameters for each of the seven steps in an octamer, the 3-step parameters, the minimum energy, the minor groove width and the RMSD).

Table 3.1: *Minimum energy parameters of the Octamer Database.*

Parameter	Description
Central step parameters	The six step parameters (twist, roll, tilt, rise, slide & shift) at the octamer central step.
All step parameters	The six step parameters (twist, roll, tilt, rise, slide & shift) at each of the seven steps of the octamer.
twist3, roll3, slide3, shift3	The four 3-step parameters (3-step twist, roll, slide & shift) at the octamer central step.
Energy	The minimum energy of the octamer.
Groove	The minor groove width, measured as the minimum phosphate-phosphate distance
RMSD	Root mean square deviation from a notional straight path through the centres of the base-pair triads.

As previously explained in Chapter 2, the base-step parameters describe the geometry between two adjacent base-pairs. The 3-step parameters measure the same rotations and translations, but in relation to the two base-pairs at the ends of a 3-step sequence (Figure 3.1). In general n-step parameters describe the geometry between the two end base-pairs of an n-step (n+1 base-pair) sequence.

Figure 3.1: *The 3-step parameters versus the single step parameters*



The minor groove width measures the minimum phosphate-phosphate distance of the minimum energy conformer minus the van der Waals' radii of the phosphate groups. Note that the major groove cannot be considered since it extends outside an octamer's length. Both grooves provide the sites at which drugs and proteins interact with DNA. The RMSD is the root mean square deviation of the actual path through the base-pair triads from a "straight" path that is aligned to the z-axis. It measures how bent a structure is.

Five percent of octamers are bistable (Gardiner et al., 2003). However for the following analysis of the parameters only the global minimum energy structures have been considered and the step parameters refer to the central step alone. Parameter distributions over the entire octamer population are described in Table 3.2. Skew is measured as the mean cube deviation (Steiner, 2000), see Equation 3.1, where N is the

population size, x_i is the ith value of the parameter, s is the standard deviation and $\bar{x}$ is the mean. Modified box plots (Weiss, 1995) of the distributions are shown in Figure 3.2 with outliers marked by crosses and determined by the inner fence boundaries.

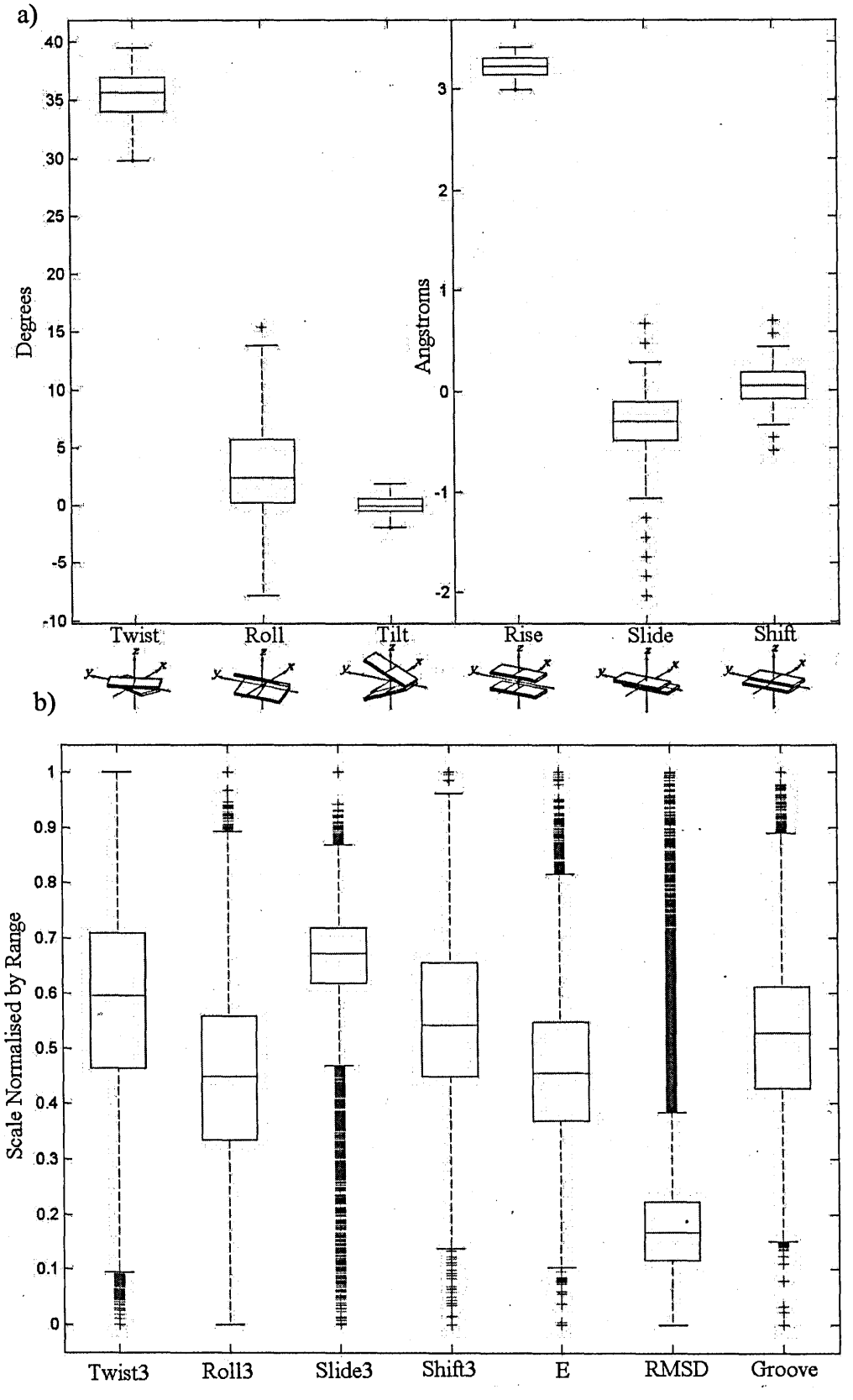
$$skew = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{s}\right)^3 \qquad \textbf{Equ. 3.1}$$

Table 3.2:. *Minimum energy structural parameter analysis. NB(Rotations are given to 1° accuracy, translations and distance are given to 0.1Å accuracy and energy 1kJmol^{-1} accuracy).*

Parameter	Minimum	Maximum	Mean	Std.Dev.	Skew
Twist	30°	39°	35°	2°	-0.58
Roll	-8°	15°	3°	4°	0.04
Tilt	-2°	2°	0°	1°	0.25
Rise	3.0Å	3.4Å	3.2Å	0.1Å	-0.37
Slide	-2.0Å	0.7Å	-0.3Å	0.3Å	-1.84
Shift	-0.6Å	0.7Å	0.1Å	0.3Å	0.32
Twist3	94°	114°	106°	4°	-0.38
Roll3	-3°	21°	8°	4°	0.11
Slide3	-6.0Å	1.7Å	-1.0Å	0.8Å	-1.99
Shift3	-0.9Å	1.2Å	0.2Å	0.3Å	-0.05
Energy	-411kJmol^{-1}	-351 kJmol^{-1}	-383 kJmol^{-1}	8kJmol^{-1}	0.28
Groove	9.4Å	13.2Å	11.4Å	0.5Å	-0.14
RMSD	0.1Å	3.0Å	0.7Å	0.4Å	2.54

Tilt, rise and shift are severely limited by the backbone (Calladine and Drew, 2002) and to a first approximation show no variation with sequence (Gardiner et al., 2003). In agreement with these observations, it can be seen from Table 3.2 and Figure 3.2a that the range of tilt is small compared to twist and roll and that rise has a negligible range of 0.4 Å in comparison to the other two translations. Roll is the single step rotation with the largest range and spread in values and twist has the largest magnitude. The analogous 3-step rotations however have similar variances to one another. Roll, roll3, shift3 and groove are normally distributed about the mean with values of skew close to zero. Both slide and slide3 have a large negative skew in their distributions. RMSD has a very positive skew of 2.54, corresponding to the fact that 90% of octamers have an RMSD of less than 1Å (Gardiner et al., 2003).

Figure 3.2: *Box plots for (a) the central 1-step parameters and (b) the central 3-step parameters, energy (E), RMSD and Groove. N.B. The parameter ranges of (b) have been normalised.*



3.2. Flexibility parameters

The flexibility parameters (Table 3.3) fall into two main categories: the force constants and the partition coefficients (see sections 3.2.1 and 3.2.2 respectively). The meaning of flexibility within the context of this work should be made clear. Flexibility of an octamer refers to the ease with which its structure can be distorted away from its global energy minimum conformation with respect to either an increase or decrease in a particular base-step parameter. Only twist and roll flexibility are considered, since these two rotations have been recognised as important in protein-DNA recognition (Koudelka et al., 1988; Rice et al., 1996).

Table 3.3: *Flexibility parameters of the Octamer Database.*

Parameter	Description
Flexibility force constants k^-_{Roll}, k^+_{Roll}, k^-_{Twist}, k^+_{Twist}	Force constants required to decrease/increase the parameter from its minimum energy value.
3-step force constants $3k^-_{Roll}$, $3k^+_{Roll}$, $3k^-_{Twist}$, $3k^+_{Twist}$	The force constants of the 3-step parameters.
Flexibility partition coefficients Q^-_{Roll}, Q^+_{Roll}, Q^-_{Twist}, Q^+_{Twist}	Partition coefficients of the single step parameters.
Total partition coefficient Q_T	Sum of the partition coefficients
3-step partition coefficients $3Q^-_{Roll}$, $3Q^+_{Roll}$, $3Q^-_{Twist}$, $3Q^+_{Twist}$	Partition coefficients of 3-step parameters.
Total 3-step partition coefficient $3Q_T$	Sum of the 3-step partition coefficients

3.2.1. The Force Constants

Both the roll energy curve and the twist energy curve of each octamer have been modelled by two Hooke's law equations, allowing highly unsymmetrical curves to be accurately represented (Gardiner et al., 2003). The energy required (E) to move an octamer from its energy minimum roll (r_{min}) to roll r is calculated by the formula below, where $x = r - r_{min}$ and k is a force constant.

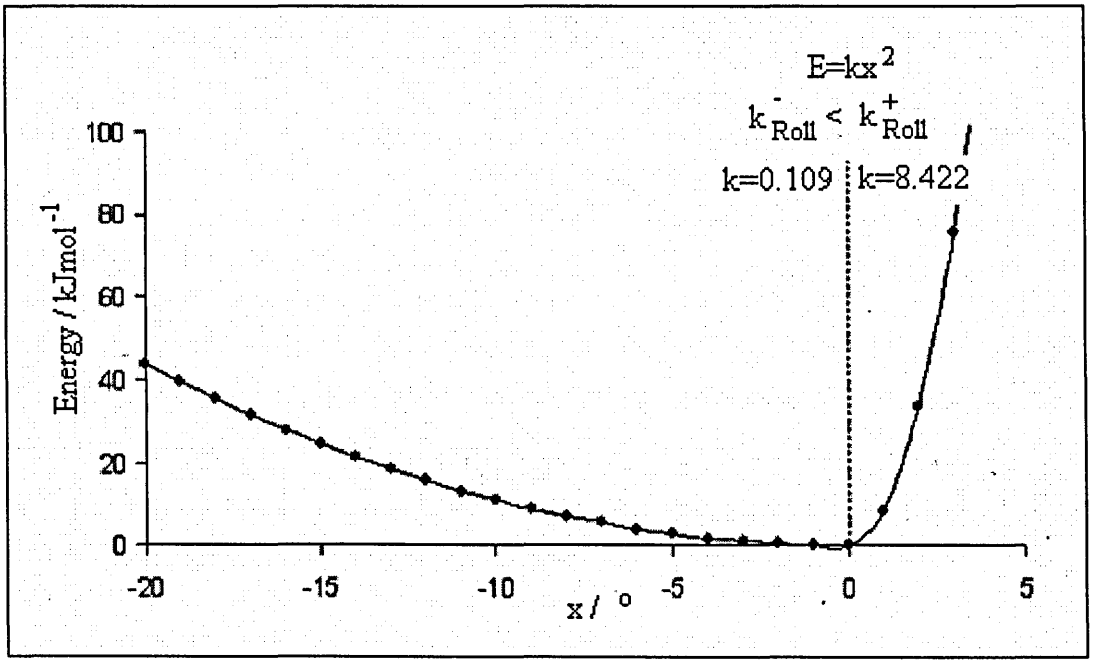
$$E = kx^2 \qquad \textbf{Equ. 3.2}$$

when $x < 0$, $k = k^-_{Roll}$

when $x > 0$, $k = k^+_{Roll}$

With analogous application to the twist energy curves, this leads to four force constants per octamer (k^-_{Roll}, k^+_{Roll}, k^-_{Twist}, and k^+_{Twist}) that describe the energy required to decrease and increase roll and twist respectively. The larger a force constant is, the steeper the curve from the minimum, meaning that more energy is required to make a rotation in that direction and the less flexible the octamer is considered to be. The modelled roll energy curve of ACCCAGCC is given in Figure 3.3. This is an extreme case of a highly unsymmetrical energy distribution, illustrating how an octamer can be flexible with respect to decrease in roll from its energy minimum structure but rigid with respect to increase in roll. Note that the value of k when x equals zero is undefined, but will always lead to zero energy.

Figure 3.3: *Roll Energy Curve of ACCCAGCC*



3.2.2. The Partition Coefficients

The partition coefficients (Q^-_{Roll}, Q^+_{Roll}, Q^-_{Twist}, Q^+_{Twist}) measure flexibility as the number of different conformations that are accessible at room temperature. Therefore the higher their values are, the more flexible the octamer. Calculations are based upon the Boltzmann distribution. First, consider roll flexibility where the Boltzmann weight, *w[x]*, is proportional to the number of conformations having a roll of r at room temperature (T=298K) with $x = r\text{-}r_{min}$ (Equation 3.3). The gas constant R (0.0083144 $kJmol^{-1}K^{-1}$) is used instead of the Boltzmann constant due to the molar energy scale. Note that *w[x]* equals one at the energy minimum, so this conformation is always populated.

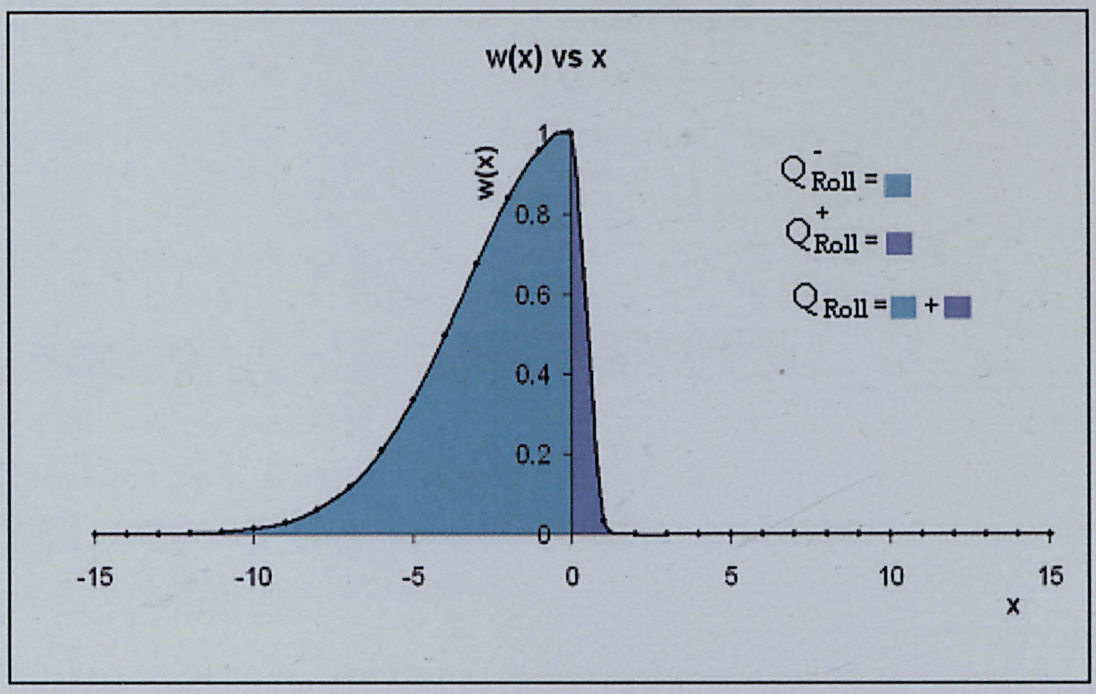
$$w[x] = \exp(-kx^2 / RT) \qquad \textbf{Equ. 3.3}$$

The total number of roll conformations available at room temperature (Q_{Roll}) is the sum of *w[x]* across all values of roll. This involves an integral over infinity due to the continuous nature of roll (Equation 3.4).

$$Q_{Roll} = \int_{-\infty}^{\infty} w[x]dx \qquad \textbf{Equ. 3.4}$$

Since k has one of two values (k^-_{Roll} and k^+_{Roll}), two separate integrals (Q^-_{Roll} and Q^+_{Roll}) must be considered and combined to calculate Q_{Roll} (Equations 3.5 a, b and c). Figure 3.4 gives a graphical illustration of the problem.

Figure 3.4: *Graphical illustration of the integrals required for calculating Q_{Roll} for ACCCAGCC.*



$$Q_{Roll} = Q^{-}_{Roll} + Q^{+}_{Roll} \qquad \textbf{Equ. 3.5a}$$

$$Q^{-}_{Roll} = \int_{-\infty}^{0} \exp(-k^{-}_{Roll} x^2 / RT)dx \qquad \textbf{Equ. 3.5b}$$

$$Q^{+}_{Roll} = \int_{0}^{\infty} \exp(-k^{+}_{Roll} x^2 / RT)dx \qquad \textbf{Equ. 3.5c}$$

Conveniently, there are exact solutions to the above definite integrals in the form shown in Equation 3.6a, where a is defined as k/RT. This results in a simple solution to the roll partition coefficient (Equation 3.6b). Note, the components (Q^{-}_{Twist} and Q^{+}_{Twist}) of the overall twist flexibility (Q_{Twist}) are calculated in an analogous way to those of Q_{Roll}.

$$\int_{0}^{\infty} \exp(-ax^2)dx = \frac{1}{2}\sqrt{\frac{\pi}{a}} \qquad \textbf{Equ. 3.6a}$$

$$Q_{Roll} = \frac{1}{2}\sqrt{\frac{RT\pi}{k^{-}_{Roll}}} + \frac{1}{2}\sqrt{\frac{RT\pi}{k^{+}_{Roll}}} \qquad \textbf{Equ. 3.6b}$$

To a first approximation the Q's are independent (Gardiner et al., 2003), meaning that they can be summed to give an overall measure of octamer flexibility, the total partition coefficient (Q_T), see Equation 3.7. The independence of the coefficients is verified in section 3.3, where parameter correlations are explored.

$$Q_T = Q^-_{Roll} + Q^+_{Roll} + Q^-_{Twist} + Q^+_{Twist} \qquad \textbf{Equ. 3.7}$$

The 3-step flexibility parameters are calculated in an identical way to above, but using the energies associated with the 3-step parameters. The minimum energy conformations can be combined with their associated flexibility parameters in order to determine the probability that an octamer will adopt a certain structure. This extension to the database is presented in Chapter 4.

Tables 3.4 and 3.4 and Figure 3.4 describe the flexibility parameter distributions. Note that the modified box plots of Figure 3.5 have had their scales normalised by their parameter range, so that degrees of skew and inner quartile ranges can be visually compared. All the force constants have positively skewed distributions, meaning that an octamer has a higher probability of being at the flexible end of a parameter's scale than at the rigid end. k^+_{Roll} is the most positively skewed 1-step force constant and has a distribution of similar shape to k^-_{Twist} (Figure 3.5). Likewise k^-_{Roll} and k^+_{Twist} have similar distribution shapes to one another. This pairing of the force constants may be due to parameter correlations, which will be explored in section 3.3. k^-_{Roll} and k^+_{Twist} have the smallest mean magnitudes (Table 3.4), meaning that on average decreasing roll and increasing twist are marginally the most flexible directions. On going from the 1-step to the 3-step force constants the positive skews are increased with a decrease in a parameter's mean and variance.

The partition coefficients distributions are more symmetric than the force constants with skews closer to zero. The 3-step partition coefficients are smoother than the 1-step with greater standard deviations. Q_{Roll} is more positively skewed than the total twist flexibility (Q_{Twist}), but this difference disappears when comparing $3Q_{Roll}$ and $3Q_{Twist}$. Increasing twist flexibility (Q^+_{Twist} or $3Q^+_{Twist}$) is the largest component of the total flexibility (approximately 30% in both cases).

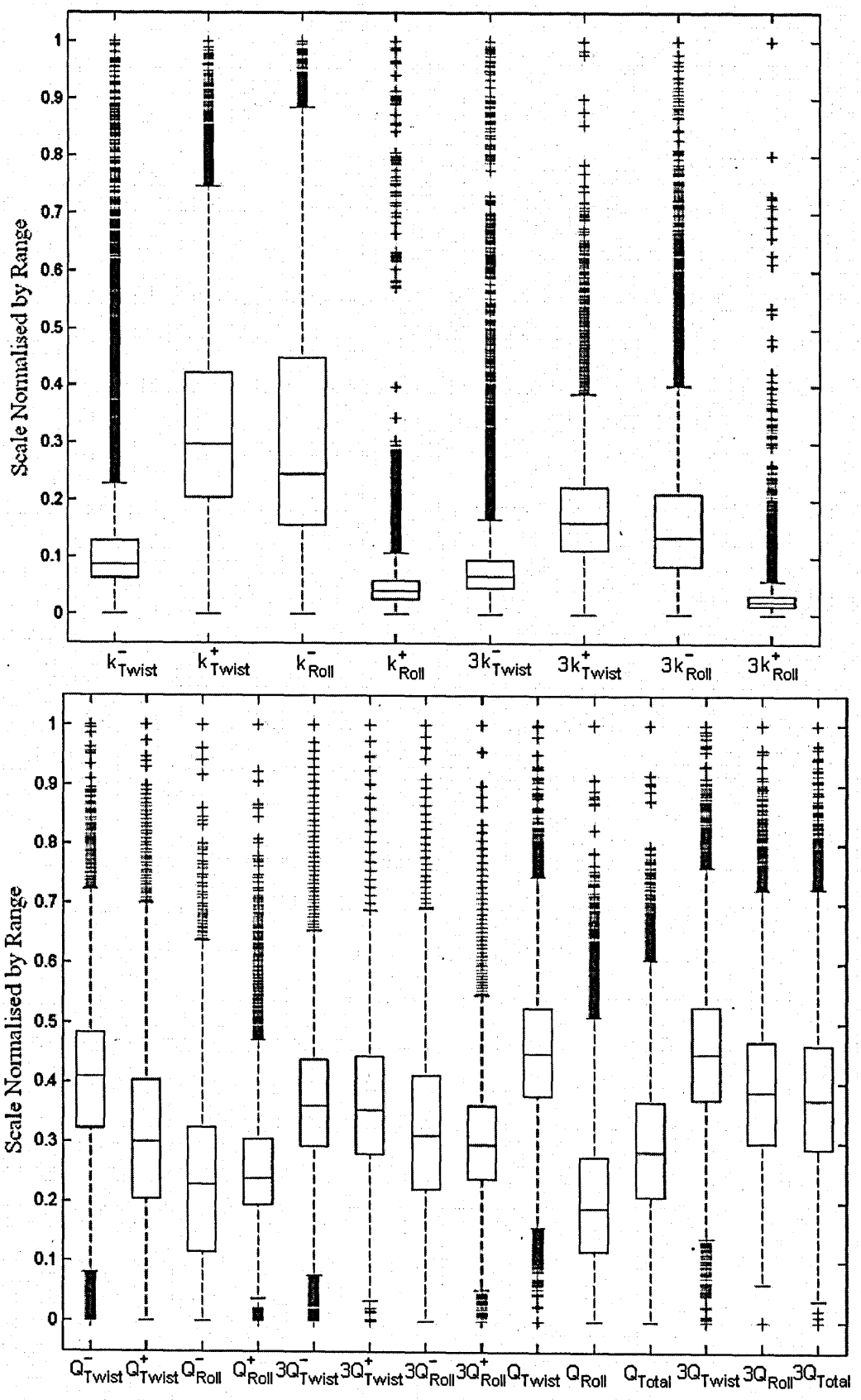
Table 3.4: *The Force Constant distribution statistics*

Parameter	Min/kJmol^{-1} degrees^{-2}	Max/kJmol^{-1} degrees^{-2}	Mean/kJmol^{-1} degrees^{-2}	Std.Dev/kJmol^{-1} degrees^{-2}	Skew
k^{-}_{Twist}	0.13	4.16	0.60	0.41	3.41
k^{+}_{Twist}	0.08	0.78	0.31	0.12	0.89
k^{-}_{Roll}	0.06	1.48	0.50	0.27	0.92
k^{+}_{Roll}	0.05	13.29	0.75	0.71	6.31
$3k^{-}_{Twist}$	0.04	2.32	0.22	0.17	5.47
$3k^{+}_{Twist}$	0.02	0.45	0.10	0.04	1.28
$3k^{-}_{Roll}$	0.03	1.01	0.20	0.12	1.75
$3k^{+}_{Roll}$	0.02	8.06	0.24	0.21	10.24

Table 3.5: *The Partition Coefficient distribution statistics*

Parameter	Minimum	Maximum	Mean	Std.Dev.	Skew
Q^{-}_{Twist}	0.69	3.96	1.99	0.43	-0.29
Q^{+}_{Twist}	1.59	5.05	2.66	0.49	0.36
Q^{-}_{Roll}	1.15	5.72	2.22	0.61	0.46
Q^{+}_{Roll}	0.38	6.33	1.90	0.61	1.01
$3Q^{-}_{Twist}$	0.92	7.29	3.26	0.74	0.26
$3Q^{+}_{Twist}$	2.09	9.24	4.70	0.86	0.50
$3Q^{-}_{Roll}$	1.39	8.09	3.53	0.91	0.39
$3Q^{+}_{Roll}$	0.49	9.45	3.22	0.88	0.70
Q_{Twist}	2.59	7.19	4.64	0.54	-0.31
Q_{Roll}	2.90	8.79	4.12	0.71	0.92
Q_T	6.11	15.07	8.77	1.02	0.66
$3Q_{Twist}$	4.14	12.53	7.96	0.97	0.33
$3Q_{Roll}$	3.46	11.95	6.76	1.05	0.33
$3Q_T$	9.50	23.16	14.71	1.73	0.45

Figure 3.5: *Flexibility Box Plots.*



3.3. Parameter Correlations

This section investigates correlations between parameters in the octamer database. Comparisons are made by calculating Spearman Rank correlation coefficients, r_s (Daly et al., 1995) and visually inspecting plots of parameter pairs. r_s measures how well the ranks are correlated to one another and has values that vary between –1 and +1. Zero means no association is present, +1 a monotonic increasing relation and –1 a monotonic decreasing relation. In the tables that follow significant correlation coefficients (where $|r_s| \geq 0.6$) are highlighted. Note that even if no strong correlation is found, a relationship may still exist that can be observed graphically.

3.3.1. Correlations between the 1-step parameters

Values of r_s for all possible pairs of the central 1-step minimum energy parameters are shown in Table 3.6. Shift and tilt (the translation and rotation of the x-axis respectively) are highly correlated (r_s of 0.95) by a positive linear relationship (Figure 3.6). This is because shift alleviates unfavourable contacts between bases that are caused by tilt. Rise and twist (the translation and rotation of the z-axis) have inversely correlated ranks, since a small rise leads to steric clashes that are minimised by increasing twist. A correlation between slide and roll is expected, since they are the translation and rotation of the y-axis. Steric clashes or electrostatic repulsions caused by change in roll may be alleviated by slide or vice versa. Surprisingly the ranks are unrelated (r_s of only 0.03), however when central step types are individually considered very strong relationships become apparent in all cases (Figure 3.7).

Table 3.6: *Spearman Rank Correlations between the 1-step parameters with values of $|r_s| \geq 0.6$ highlighted*

	Twist	Roll	Tilt	Rise	Slide	Shift
Twist	1					
Roll	-0.36	1				
Tilt	-0.22	***-0.64***	1			
Rise	***-0.70***	-0.26	0.58	1		
Slide	0.19	0.03	0.02	0.05	1	
Shift	-0.15	***-0.64***	***0.95***	0.49	0.03	1

Figure 3.6: *Shift versus Tilt.*

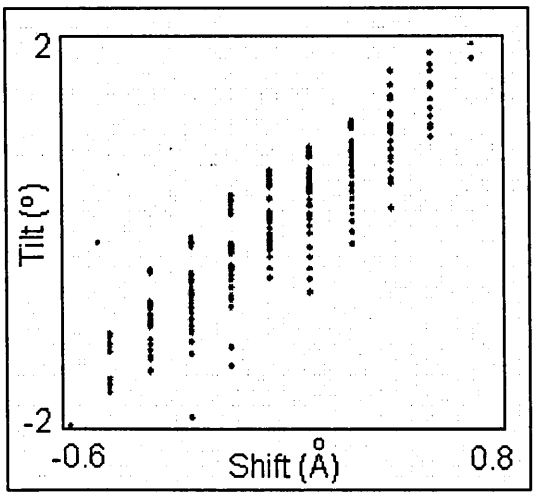
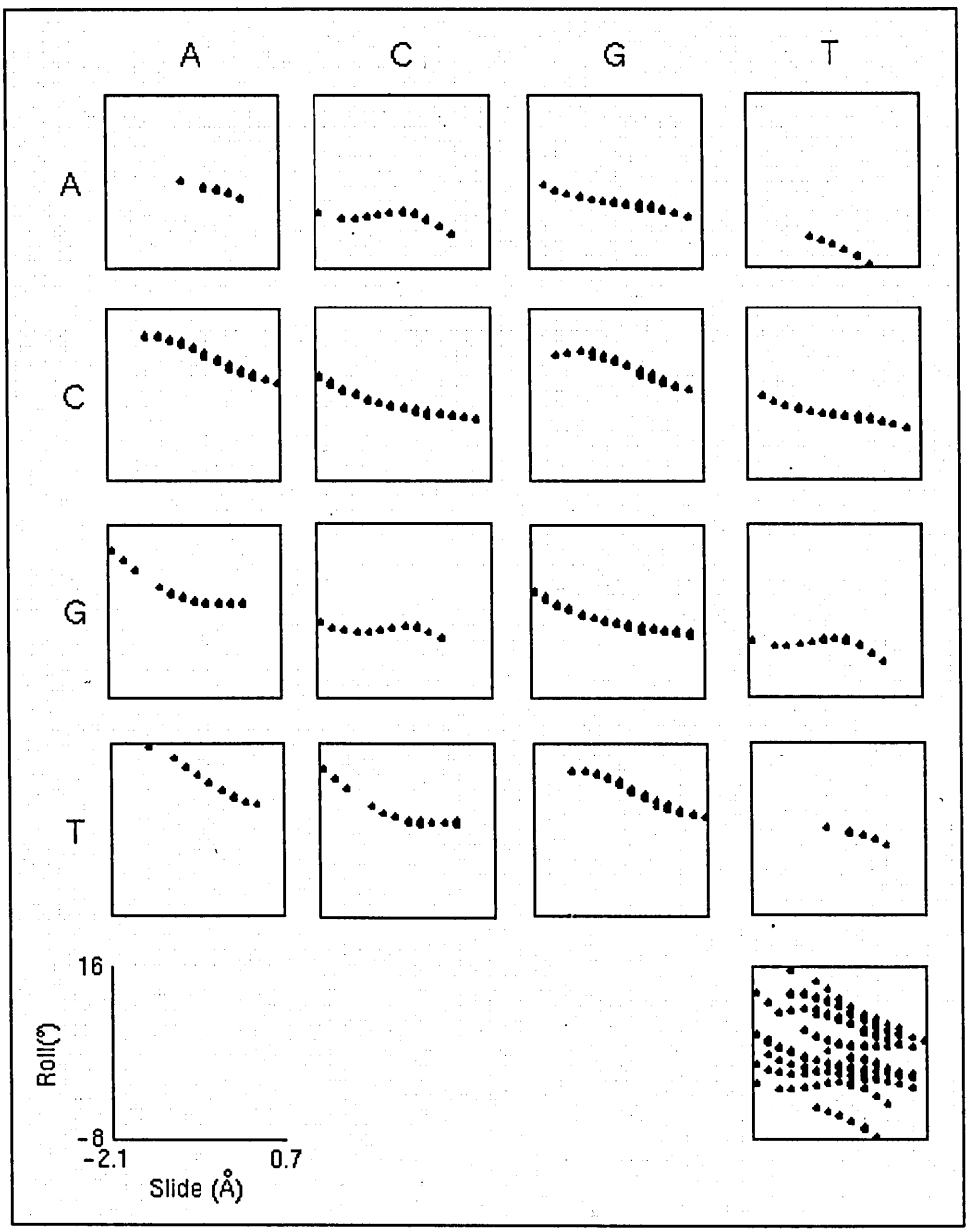


Figure 3.7: *Slide versus roll for different central steps. All octamers are shown in bottom right plot*



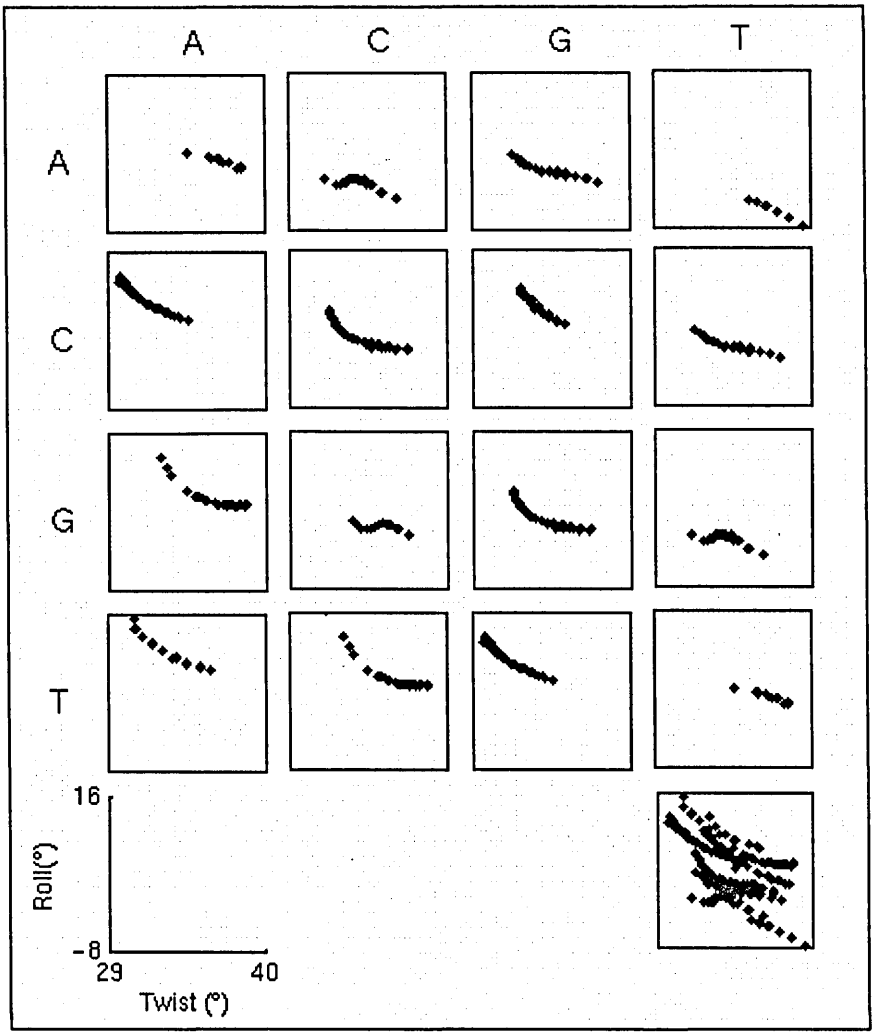
In eight out of the ten central step types, an inverse relationship can be seen between slide and roll (Figure 3.7). Notice that both the GC and AC/GT plot are wave-like in shape. The structural reasons for this are unknown. They are both purine-pyrimidine steps, but more specifically they are the only two guanine-pyrimidine steps. Perhaps when certain parts of the guanine base slide over the other purine base, a steric clash or electrostatic repulsion occurs that causes a temporary change in the direction of roll.

The shift-roll relationship is very similar to the tilt-roll relationship, due to the strong correlation between shift and tilt. High minimum energy roll tends to favour low shift and tilt. Previous research has suggested that roll and twist are anti-correlated with the exact linear relationship varying with the central step type (Gorin et al., 1995). The central step roll-twist graphs of Figure 3.8 look very similar to their analogous central step roll-slide graphs of Figure 3.7. This suggests that there is a very strong positive correlation between slide and twist for each central step type. The linear nature of this relationship is confirmed by the high squared Pearson correlation coefficients (r^2) of Table 3.7. Coefficients are also included for the roll-twist and roll-slide relationships. In general, roll, twist and slide are all highly coupled to one another. Notice the low roll-twist r^2 values for the TC/GA central step (Table 3.7), signifying non-linearity. The curvi-linear nature of these relationships is clearly seen in Figures 3.7 and 3.8.

Table 3.7: *Squared Pearson correlation coefficients of slide-twist, roll-twist & roll-slide correlations.*

Step type	Slide-twist	Roll-twist	Roll-slide
TG/CA	0.92	0.95	0.98
TC/GA	0.96	0.40	0.35
TA/TA	0.98	0.95	0.98
CG/CG	0.89	0.94	0.92
GC/GC	0.96	0.16	0.05
GG/CC	0.97	0.86	0.94
AT/AT	0.99	0.99	0.98
AG/CT	0.98	0.86	0.90
AC/GT	0.91	0.73	0.57
AA/TT	0.99	0.98	0.98
All octamers	0.08	0.31	0.00

Figure 3.8: *Twist versus roll for different central steps. All octamers are shown in bottom right plot.*



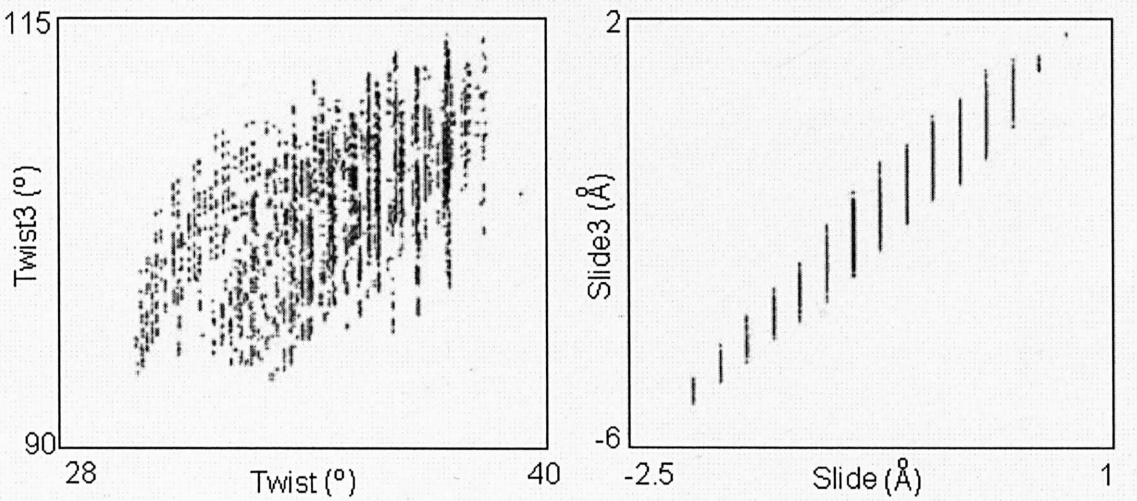
3.3.2. Correlations between the 3-step parameters

The roll-shift correlation disappears on going from the central single step parameters to the central 3-step parameters. There is no strong monotonic relationship between any of the 3-step parameters (Table 3.8). Slide has a tendency to be similar along a sequence (Packer et al., 2000b), hence the strong correlation between slide and slide3 (Table 3.8 and Figure 3.9). The plot of twist versus twist3 in Figure 3.9 emphasises that the 3-step parameters (apart from slide) contain information very different from their single step counterpart, therefore validating their use.

Table 3.8: *Spearman Rank Correlations between the 3-step parameters with the most significant highlighted (slide3-slide).*

	Twist3	Roll3	Slide3	Shift3	1-Step Equivalent
Twist3	1				0.59
Roll3	-0.49	1			0.26
Slide3	0.57	-0.31	1		**0.85**
Shift3	-0.02	-0.14	0.15	1	-0.17

Figure 3.9: *1-step to 3-step correlations*



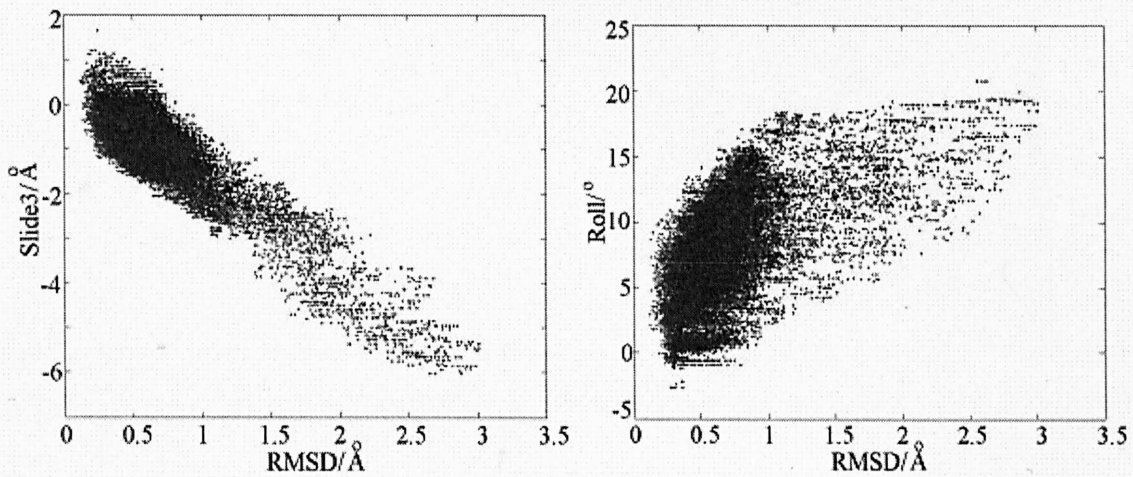
3.3.3. Energy, groove and RMSD correlations

No significant energy correlations exist with all magnitudes of r_s being less than 0.39 (Table 3.9). Both groove and RMSD are inversely correlated to twist3 with an r_s of –0.60. An inverse relationship between groove and twist is understandable, because untwisting opens the groove (Gorin et al., 1995). RMSD is inversely correlated to both slide3 (Figure 3.10) and slide, r_s of –0.70 and –0.61 respectively. The RMSD versus roll plot is included in Figure 3.10 since it shows a tendency of roll to be high when RMSD is high, meaning that roll is an important degree of freedom in bent structures.

Table 3.9: *Spearman Rank Coefficients for energy, groove and RMSD correlations with the most significant highlighted (RMSD-Slide3, RMSD-Slide, RMSD-Twist3, Groove-Twist3)*

	Energy	Groove	RMSD
Twist	-0.17	-0.41	-0.37
Roll	-0.04	0.01	0.25
Tilt	0.17	0.18	-0.09
Rise	0.21	0.36	0.00
Slide	0.10	-0.11	**-0.61**
Shift	0.08	0.11	-0.08
Twist3	-0.21	**-0.60**	**-0.60**
Roll3	-0.02	0.35	0.59
Slide3	0.12	-0.14	**-0.70**
Shift3	0.06	-0.08	-0.04
Energy	1	0.39	-0.15
Groove	0.39	1	0.24
RMSD	-0.15	0.24	1

Figure 3.10: *Most significant RMSD and Groove correlations*



3.3.4. Force constants & Partition Coefficients

Flexibility in the increasing twist direction is positively correlated to flexibility in decreasing roll with a 1-step r_s of 0.61 and a 3-step r_s of 0.66 (Tables 3.10 and 3.11). Note that the partition coefficient pairs will have the same r_s values as the analogous force constant pairs, due to the derivation of one from the other. The Pearson

correlation coefficients will be different however, since although the ordering of the points within a graph will be the same, their dispersion will be different (Figures 3.11a and b). k^+_{roll} is only high when k^-_{roll} is low (Figure 3.11c) and an octamer never appears highly rigid with respect to both decrease in roll and decrease in twist (Figure 3.11d), meaning that an octamer is always able to either untwist or decrease roll to a certain extent, which are both mechanisms of relieving clashes in the major groove. It can be confirmed that the partition coefficients are independent to a first approximation, since (apart from Q^+_{twist} and Q^-_{roll}) all pairs have an r_s of less than 0.5. This validates the use of Q_{Twist}, Q_{Roll} and Q_T as descriptors.

Table 3.10: *Single step force constant correlations with the most significant highlighted (k^+_{twist}-k^-_{roll})*

	k^-_{Twist}	k^+_{Twist}	k^-_{Roll}	k^+_{Roll}
k^-_{Twist}	1			
k^+_{Twist}	-0.35	1		
k^-_{Roll}	-0.48	**0.61**	1	
k^+_{Roll}	0.43	-0.19	-0.37	1

Table 3.11: *3-step force constant correlations*

	$3k^-_{Twist}$	$3k^+_{Twist}$	$3k^-_{Roll}$	$3k^+_{Roll}$	1-Step Equivalent
$3k^-_{Twist}$	1				**0.61**
$3k^+_{Twist}$	-0.26	1			**0.65**
$3k^-_{Roll}$	-0.45	**0.66**	1		**0.73**
$3k^+_{Roll}$	-0.54	-0.12	-0.30	1	0.59

3.3.5. Correlations between flexibility and minimum energy

Spearman rank correlation coefficients between a selection of flexibility parameters and minimum energy parameters are given in Table 3.12. The only significant correlation found was an inverse monotonic relation between minimum energy 3-step twist and flexibility in the increasing twist direction. A low r_s of 0.09 between Q_T and RMSD confirms that how bent an octamer is in its minimum energy conformation is unrelated to its overall flexibility (Gardiner et al., 2003).

Figure 3.11: *Some force constant and partition coefficient correlations*

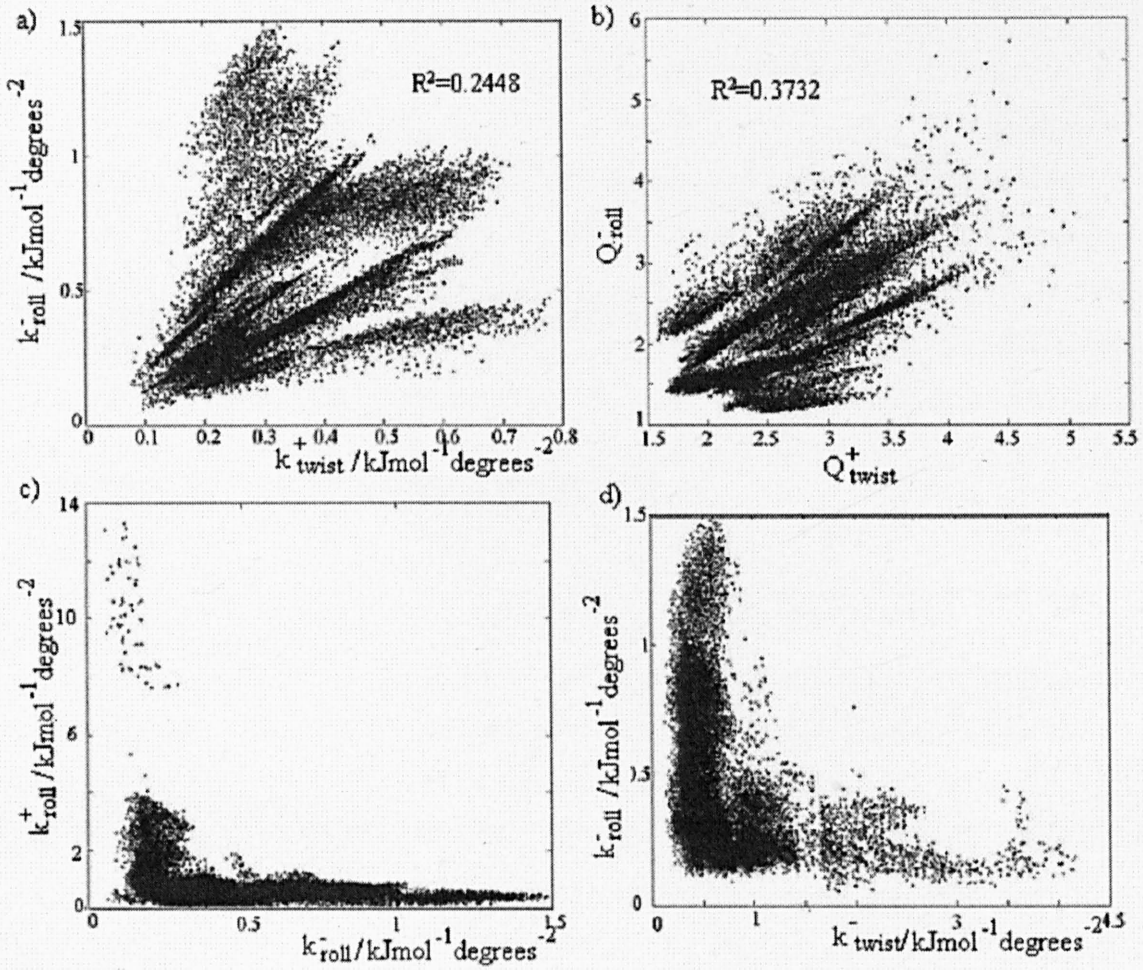


Table 3.12: *Correlations between flexibility and minimum energy*

	k^-_{Twist}	k^+_{Twist}	Q^-_{Twist}	Q^+_{Twist}	Q_{Twist}	
Twist	-0.5553	0.4905	0.5553	-0.4905	-0.0689	
	k^-_{Roll}	k^+_{Roll}	Q^-_{Roll}	Q^+_{Roll}	Q_{Roll}	
Roll	-0.1576	-0.1874	0.1576	0.1874	0.2781	
	$3k^-_{Twist}$	$3k^+_{Twist}$	$3Q^-_{Twist}$	$3Q^+_{Twist}$	$3Q_{Twist}$	
Twist3	-0.3973	**0.6938**	0.3973	**-0.6938**	-0.3172	
	$3k^-_{Roll}$	$3k^+_{Roll}$	$3Q^-_{Roll}$	$3Q^+_{Roll}$	$3Q_{Roll}$	
Roll3	-0.3687	0.0679	0.3687	-0.0679	0.2768	
	QTwist	QRoll	QTotal	3QTwist	3QRoll	3QTotal
E	0.2750	0.1796	0.2436	0.5276	0.3978	0.5336
Groove	0.2730	0.2410	0.2821	0.3984	0.4010	0.4666
RMSD	-0.1111	0.2236	0.0853	-0.1382	0.2479	0.0770

3.4. Conclusions

A database that describes the minimum energy structure and flexibility of all DNA octamers has been successfully produced (Gardiner et al., 2003). The parameters tilt, rise and shift show little variation due to the backbone constraints. Slide is negatively skewed with a tendency to be similar along a sequence, hence its strong positive correlation to slide3. Twist and roll are both important rotations for protein recognition that are anti-correlated. Their exact correlation varies with the central-step and is found to be identical to the corresponding slide and roll relationship. RMSD is positively skewed with most octamers having values less than 1 Å. Roll tends to be high when RMSD is high, meaning that it is important degree of freedom in bent structures. How bent an octamer is has no relation to its flexibility.

On average, increasing twist is the most favoured direction in flexibility, followed by decreasing roll. Increasing twist flexibility is positively correlated to that of decreasing roll. An octamer never appears to be highly rigid with respect to both decrease in roll and decrease in twist, therefore there is always some degree of flexibility for widening the major groove, a common way by which proteins bind to DNA (Brandon and Tooze, 1991). The following chapter combines an octamer's minimum energy conformation with its flexibility, in order to determine its structural probabilities.

Chapter 4:
Database Extension – Structural Probabilities

The minimum energy conformation and flexibility of an octamer can be combined to calculate structural probabilities in terms of roll or twist. This fusion of database parameters is explored and offers a useful alternative way of comparing DNA sequences to one another. It will enable estimates of the likelihood that two sequences will have the same structure or that a sequence will adopt a certain binding motif. The probabilities must be calculated using a numerical integration technique due to the continuous nature of the structural parameters. A Rectangular Approximation algorithm has been implemented for this purpose and is found to give accurate results.

4.1. Calculating the probabilities

The probability that an octamer has an exact value of roll (r) is mathematically impossible to calculate, since roll is a continuous variable. This means that the number of possible outcomes is infinite with the probability of one particular outcome occurring being undefined. Therefore, only the probability that an octamer will have a roll within a defined range *a* to *b*, $P[a \leq r \leq b]$, can be considered. See Equation 4.1, where *w[x]* is the Boltzmann weight and Q_{Roll} is the roll partition coefficient (Chapter 3 section 3.2).

$$P[a \leq r \leq b] = \frac{\int_{b}^{a} w[x]}{Q_{Roll}} \qquad \textbf{Equ. 4.1}$$

Unlike the calculation of Q_{Roll} discussed in Chapter 3, there is no standard solution to the integration of *w[x]* over defined ranges and numerical integration must be used in order to approximate areas under a curve. The simplest method is the Rectangular Approximation where the area is represented by rectangles of fixed widths (Figure 4.1). The height of a rectangle is defined by the value of the function at the midpoint of the rectangle width. The Trapezoidal rule and the Simpson's Rule are two

other commonly used approximations that also use uniformly spaced ordinates (Figure 4.1). They are known as the Newton-Cotes quadratures. The Trapezoidal rule uses trapezia rather than rectangles and Simpson's rule uses parabolas. Formulae for the approximations are given in Equations 4.2a, b and c. It can be seen that all three methods are relatively straightforward to calculate and not computationally demanding.

Rectangular:

$$\int_a^b f(x)dx \approx h \sum_{0 \le i \le n} f\left(\frac{x_i + x_{i+1}}{2}\right) \qquad \textbf{Equ. 4.2a}$$

where $h=(b-a)/n$ for n rectangles

Trapezoidal:

$$\int_a^b f(x)dx \approx h\left[\frac{1}{2}f(x_0) + f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + \frac{1}{2}f(x_n)\right] \qquad \textbf{Equ. 4.2b}$$

where $h=(b-a)/n$ for n trapezia

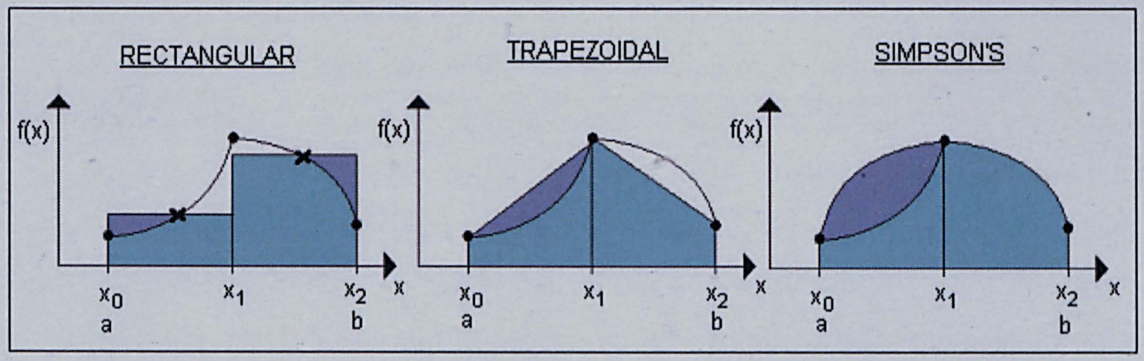
Simpson's:

$$\int_a^b f(x)dx \approx \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \ldots + 4f(x_{2n-1}) + f(x_{2n})\right]$$

Equ. 4.2c

where $h=(b-a)/2n$ for n parabola

Figure 4.1: *Methods of Numerical Integration. Illustrations of the Rectangular, Trapezoidal and Simpson's rule for estimating the integration of a theoretical function, f(x), using three ordinates x_0, x_1 and x_2. Shaded areas show the estimations with the darker shading referring to regions above the curve. Crosses in the Rectangular illustration show the midpoint heights.*



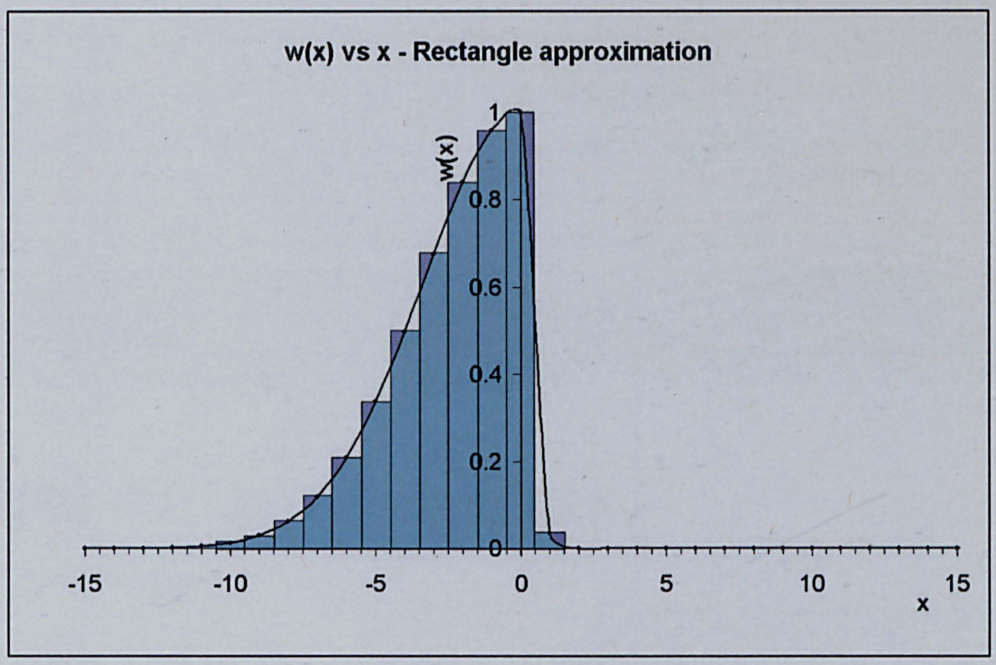
The use of rectangles can be more accurate than trapezia, because a rectangle's tendency to go partly over and under a curve may cancel out the error in area estimation, whereas a trapezium tends to completely under-estimate or over-estimate the area (Sedgewick, 1988). This is illustrated in Figure 4.1. The over-estimation between x_0 and x_1 of the Simpson's approximation in Figure 4.1 illustrates how parabolas of smaller widths (i.e. one from x_0 to x_1 and a second from x_1 to x_2) would allow a much better estimation. This inaccuracy may also be avoided by using a Gaussian quadrature. Gaussian quadratures, unlike Newton-Cotes quadratures, have unevenly spaced ordinates chosen to optimise the area estimation. Further information on Gaussian and Newton-Cotes quadratures can be found in the literature (Sedgewick, 1988; Acton, 1990; Borse, 1997; Steiner, 2000).

Acton suggests that given the availability of computers to automate processes using the simplest technique repeatedly is often the most efficient and effective approach:

> "Why not count squares – provided we have an automated computer to do it?"
> (Acton, 1990)

This opinion supports the decision to use the Rectangular approximation. An algorithm has been written that starts estimating the desired area by splitting it into a specified starting number of rectangles. Estimations are then repeated with doubling of the number of rectangles until the solution has converged to a required number of decimal places. Since the function being numerically integrated, *w[x]*, depends on whether x is positive or negative (whether roll/twist is being decreased or increased from its energy minimum) it is important to treat areas that cross an x of zero with caution. Figure 4.2 illustrates an extreme case of a highly unsymmetrical distribution (the roll of octamer ACCCAGCC), showing how a rectangle that crosses the energy minimum may lead to inaccurate predictions. For this reason calculations will treat areas that cross zero as the sum of two separate area estimations, allowing convergence to occur with a differing number of rectangles either side of the energy minimum. Finally once the desired integral has been estimated it can be converted into its associated probability with division by the partition coefficient.

Figure 4.2: *The rectangle approximation for the roll of ACCCAGCC*



The convergence threshold was set to 0.00001 (5 decimal place accuracy) and the number of starting rectangles was initially set to two. However it was discovered that under certain circumstances the area estimation algorithm converged prematurely. For example the probability that CGGTATAC has a roll between -10^{o} and $+20^{o}$ is approximately one, but when considering the broader range of -100^{o} to $+100^{o}$ the probability estimation dramatically decreases to 0.4525. This severe error and reduction in probability is caused by a drastic under-estimation of the area to the left of r_{min} (a decrease from 3.15 to $6.01X10^{-6}$ degrees). Convergence has occurred much too early at only four rectangles, each with an approximate width of 28^{o}. A simple but effective solution to this problem is used. Instead of setting the starting number of rectangles to two regardless of the roll range size, a starting number that corresponds to rectangle widths of one degree is used.

Now let's return to the highly unsymmetrical roll probability distribution of ACCCAGCC (Figure 4.2). Differing numbers of optimal rectangles either side of ACCCAGCC's minimum energy roll structure (r_{min}) have been found (Table 4.1). This justifies the use of a probability calculation procedure that estimates areas either side of the energy minimum independently. Note that ACCCAGCC has a r_{min} of 11.90^{o}, k^{-}_{roll} of 0.109 $kJmol^{-1}degrees^{-2}$ and k^{+}_{roll} of 8.422 $kJmol^{-1}degrees^{-2}$. The range -10^{o} to $+20^{o}$ is included in Table 4.1, since this is the roll variation in usual DNA structures (Calladine

and Drew, 2002). As expected, the respective probability is one. With the exception of $r_{min} \pm 1^o$, the number of rectangles is greater below r_{min} than above. The optimal number of rectangles depends on the gradient magnitudes and the pace with which they are changing, defined by the force constant and distance from minimum energy roll.

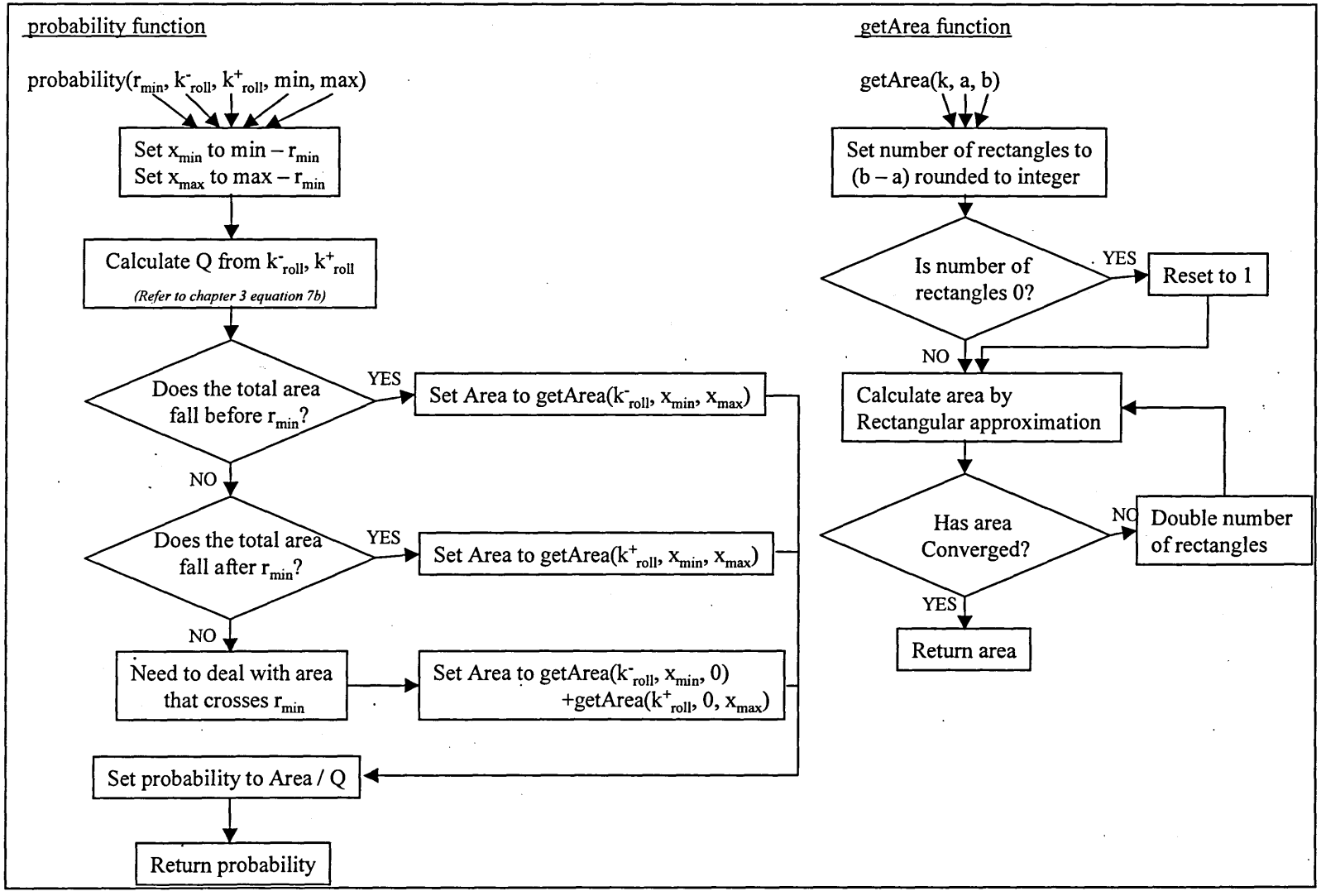
Table 4.1: *ACCCAGCC probability calculations around its energy minimum roll (r_{min})*

Probability range	No of rectangles before r_{min}	No of rectangles after r_{min}	Probability
$r_{min} \pm 1^o$	128	128	0.311
$r_{min} \pm 2^o$	256	16	0.503
$r_{min} \pm 3^o$	384	24	0.665
$r_{min} \pm 4^o$	512	32	0.788
$r_{min} \pm 5^o$	640	40	0.876
$r_{min} \pm 10^o$	320	80	0.997
-10^o to $+20^o$	320	160	1.000

It can be confirmed that identical probabilities to those of Table 4.1 are obtained when using the more sophisticated, but also more computationally demanding, Labatto quadrature to estimate the required integrals. The Labotta quadrature is a recursive adaptive gaussian procedure that is available via the 'quadl' function in Matlab (Gander and Gautschi, 2000).

Flowcharts showing details of the final structural probability algorithm and its implementation are given in Figure 4.3. Two functions are shown, the probability function and the getArea function. The former makes calls to the later. A third function also exists (RectangularApproximation) but has not been included since its structure is covered in adequate detail by Equation 4.2a. Program flow is initiated by a call to the probability function. Five input arguments are needed: r_{min}, k^-_{roll}, k^+_{roll}, the lower roll probability limit (min) and the upper roll probability limit (max). The main part of this function is to determine what areas need to be calculated. Three situations exist: the whole area occurs either before or after r_{min} or it crosses r_{min}. As discussed above, if it crossed r_{min} then the area is split into two, resulting in two calls to the getArea function. Once called, the getArea function calculates an area using a determined number of starting rectangles and an iterative procedure until the convergence threshold is reached.

Figure 4.3: *The probability and getArea functions of the Rectangular Approximation algorithm*



Although the majority of DNA structures have a single-step roll between -10^{o} and $+20^{o}$ it was found that in order to cover the structural space available to the octamer population an extended range of -20^{o} to $+30^{o}$ must be used. The 3-step roll and analogous twist dimensions are shown in Table 4.2. These dimensions will need to be considered when comparing the structures of two octamers, the topic of the next section.

Table 4.2: *The structural Roll/Twist space*

Dimension	Minimum / ^{o}	Maximum / ^{o}
Roll	-20	30
Roll3	-20	40
Twist	15	55
Twist3	75	135

4.2. Structural Similarity

The probability that an octamer will have a particular range of roll can now be calculated (section 4.1). This leads to a further question of how to calculate the probability that two octamers will have the same roll, their structural similarity. It is important to note that although all examples so far have been about roll the same principles are used to calculate twist probabilities and similarities.

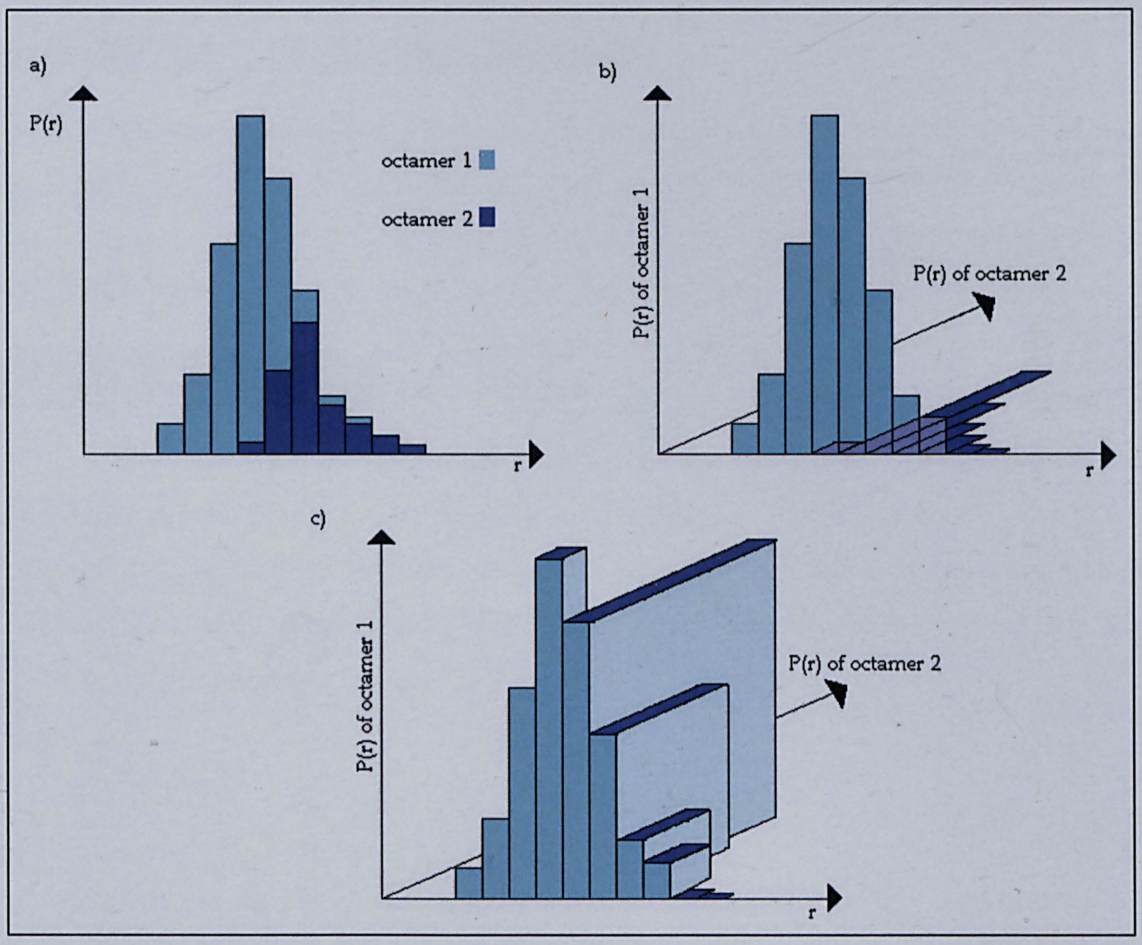
Consider two octamers X and Y. X has a roll of X_{roll} and Y has a roll of Y_{roll}. The probability that X and Y have identical roll structures, $P(X_{roll}=Y_{roll})$, is estimated as the sum of probability products over roll's structural space, see Equation 4.3. The structural space, although continuous, is considered in one-degree segments. Hence Equation 4.3 calculates the probability that both octamers have a roll in the first one-degree bin plus the probability they both have a roll in the next bin and so on until the entire roll dimension has been considered.

$$P(X_{roll} = Y_{roll}) = \sum_{i=-20}^{+30} P(X_{roll} = r_i \pm 0.5) P(Y_{roll} = r_i \pm 0.5) \qquad \textbf{Equ. 4.3}$$

This procedure is graphically illustrated as a volume in Figure 4.4. The quantized probability distributions of two hypothetical octamers are shown in the same plane in Figure 4.4a, then at right angles to one another in Figure 4.4b. Three dimensional blocks can then be drawn to join up the distributions (Figure 4.4c), the total volume of which estimates $P(X_{roll}=Y_{roll})$.

Figure 4.4: *Estimating $P(X_{roll}=Y_{roll})$*

(a) Separate probability distributions, (b) rearrange distributions to be at right angles to one another, (c) then calculate volume of blocks that join the distributions together.



The structural similarity of two octamers X and Y, S(X,Y), will have a value between zero and one, one meaning identical. This allows comparisons between different structural similarity measures to be made, see the next section. The probability that two octamers will have the same structure (either with respect to roll or twist) will never be equal to one, since a DNA sequence's structure is not static. Even when

considering a pair of identical octamers the chances that they will both be in the same configuration at the same moment in time is far from certain. In fact the average probability that a pair of identical octamers will have the same roll structure is only 0.174 or that they will have the same twist structure is only 0.153. The highest pairwise roll probability (0.238) is between AGAGAATT and itself or its structural equivalent (AATTCTCT), since it is the octamer most rigid to changes in roll with the lowest value of Q_{Roll}. It may seem alarming that structural probabilities are so low between identical octamers, but they clearly reflect the importance of dynamics in DNA structure, a long recognised characteristic (Levitt, 1983). Note however that the probabilities only consider octamers in isolation and do not account for any structural constraints that may be placed upon them by surrounding base-pairs or other environmental factors

It can be more probable for two different octamers to have the same central step geometry than two identical octamers. For example AGGTAGCC is more likely to have the same value of roll compared with AGGTAGAG than with another molecule of itself. This is because AGGTAGCC is extremely flexible by roll and has a very similar minimum energy roll to the less flexible AGGTAGAG.

Two normalisation techniques are presented for the conversion of the probability measures to symmetric non-directional similarities, i.e. S(X,Y) is equal to S(Y,X). Method 1 always gives an octamer a similarity of 1 with itself, whereas method 2 differentiates between the level of similarity between one identical pair and another, i.e. S(X,X) does not necessarily equal S(Y,Y).

- *Method 1:*

$S(X,Y) = S(Y,X)$

$S(X,X) = 1$

$$S(X,Y) = \frac{2P(X=Y)}{P(X=X)+P(Y=Y)} \qquad \textbf{Equ. 4.4a}$$

- *Method 2:*

$S(X,Y) = S(Y,X)$

Differentiates whether $P(X=X) > P(Y=Y)$

$$S(X,Y) = \frac{P(X=Y)}{MAX}, \qquad \textbf{Equ. 4.4b}$$

where MAX is the maximum pairwise probability over the entire octamer population (i.e. 0.238 for the conversion of the roll probabilities). In other words, the similarities values are all relative to the most similar octamer pair (that is the similarity of the most rigid octamer to itself).

N.B.(Do not confuse MAX with $max\{P(X=X), P(Y=Y)\}$).

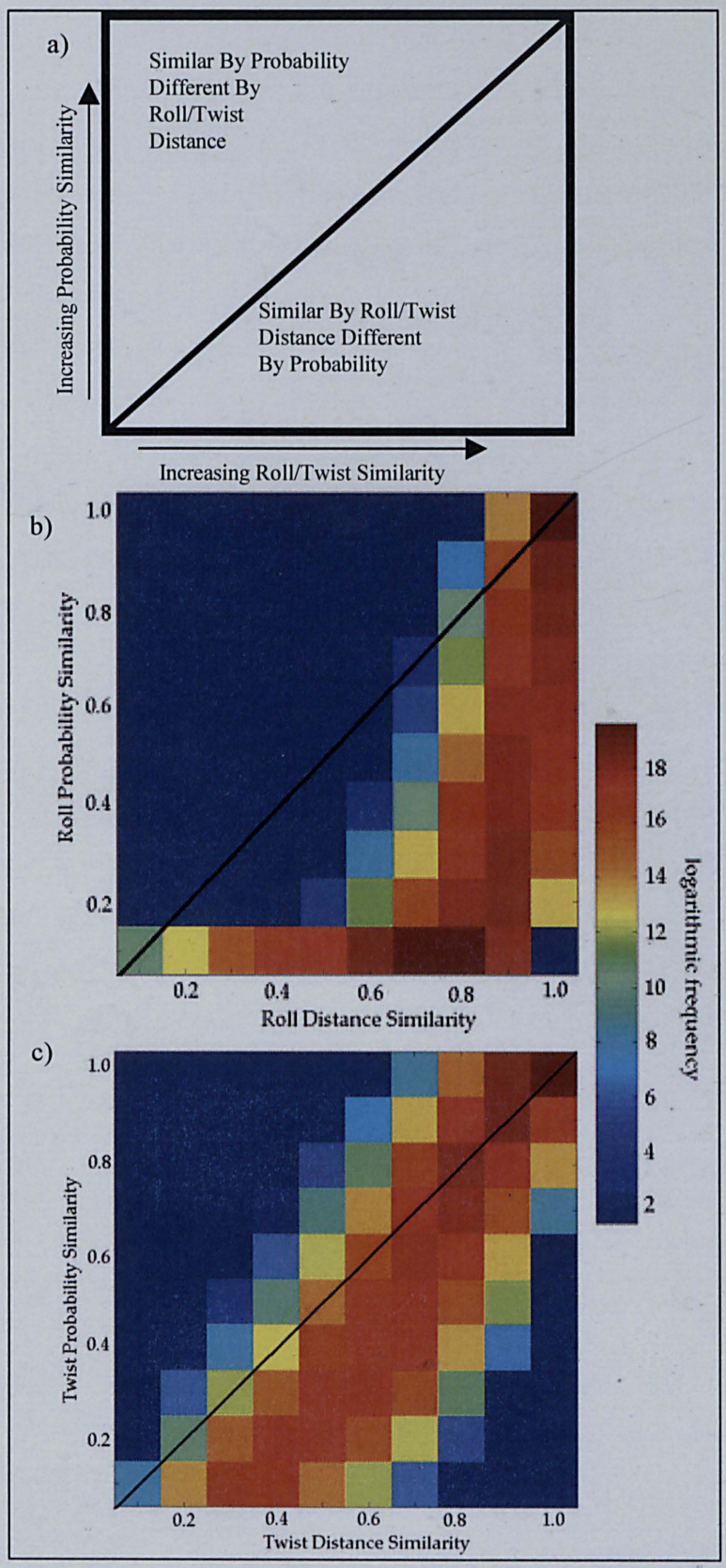
4.3. Comparison to minimum energy structure distances

A similarity measure based upon structural probabilities will be of no use if it is found to give similarity scores identical to those based upon the minimum energy structures. Equation 4.5 defines the minimum energy structure similarity between octamers X and Y when considering a single parameter p. $D_p(X,Y)$ is the distance between X and Y with respect to p and r is the parameter's population range.

$$S(X,Y) = 1 - \frac{D_p(X,Y)}{r} \qquad \textbf{Equ. 4.5}$$

It is a daunting task to look at all possible pairs of octamers, therefore similarities have been binned and frequency matrices drawn for both the roll and twist comparisons (Figure 4.5). The measures introduced in Equations 4.4a and 4.5 are used. If the similarity measures were equivalent, the use of both would be redundant and all the scores would lie on the diagonal of the matrix. The area above the diagonal corresponds to an octamer pair being more similar by their structural probability than by their minimum energy structure and vice versa for the area below the diagonal (Figure 4.5a).

Figure 4.5: *Structural probability and minimum energy structure similarity comparisons for (b) 1-step roll and (c) 1-step twist.*



In the roll matrix (Figure 4.5b) the majority of octamer pairs lie beneath the diagonal, meaning that they are less similar by their structural roll probabilities than by their r_{min} values. 49% of the octamer pairs have a roll distance similarity greater than 0.7 and a roll probability similarity of less than 0.4. This reflects the fact that if a pair of octamers have identical or near identical minimum energy structures they can still have very different structural probabilities, due to differences in their flexibility. On the other hand, when the roll probabilities are high it is impossible for the r_{min} similarity to be low, due to the relationship between the r_{min} values and roll probability. Hence the absence of octamers in the top left hand corner of both plots and the asymmetric nature of the matrices. In the twist matrix (Figure 4.5c) the majority of octamer pairs are also below the diagonal, though to a far lesser extent than in the roll matrix. Clearly different information is contained in the novel probability similarities in comparison to their minimum energy counterparts, justifying their use as alternative descriptors for sequence comparison methods.

Figure 4.6a shows the roll probability distributions of an octamer pair that are similar by both their minimum energy roll and their roll probabilities. The distributions have almost identical shapes with a very large overlap. The similarity values are given in Table 4.3 along with each octamer's r_{min}, k^{-}_{roll} and k^{+}_{roll}. Figure 4.6b shows the roll probability distributions of another octamer pair that are again similar by their minimum energy roll, but are very different by their roll probabilities. GGGGAGTC (shown in black) is highly flexible with respect to increasing roll (Table 4.3) making its structural roll tendencies very different from AAAAAACA.

Figure 4.6: *Boltzmanns weight, w(x), versus roll for (a) an octamer pair (AAAAAATC and TGGGCATA) similar by both their minimum energy roll and roll probabilities and (b) an octamer pair (AAAAAACA and GGGGAGTC) similar by their minimum energy roll but very different by their roll probabilities.*

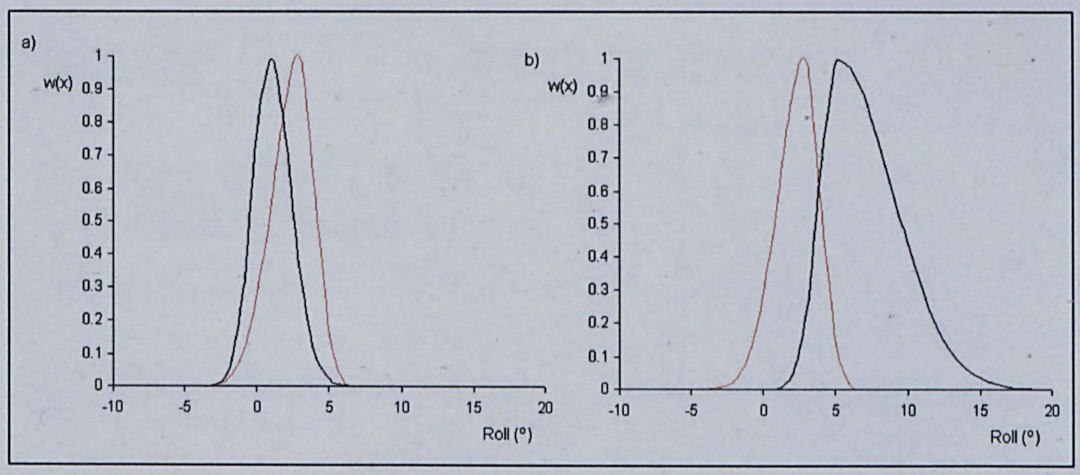


Table 4.3: *The similarities between the octamer pairs shown in Figure 4.6 along with their minimum energy roll and roll flexibility values.*

Octamer	r_{min} (degrees)	k^{-}_{roll} ($kJmol^{-1}$ $degrees^{-2}$)	k^{+}_{roll} ($kJmol^{-1}$ $degrees^{-2}$)	Roll distance similarity	Roll probability similarity
AAAAAATC	3	0.37	1.13	0.9	0.8
TGGGCATA	1	0.95	0.57		
AAAAAACA	3	0.37	1.13	0.9	0.2
GGGGAGTC	5	0.77	0.09		

4.4. Conclusions

The Rectangular Approximation algorithm is a fast and efficient iterative procedure with a 5 decimal place convergence threshold that calculates the probability that an octamer will adopt a particular roll or twist structure. It has been confirmed that combining the minimum energy conformation and flexibility of an octamer to evaluate its structural tendencies does provide a novel way of comparing sequences by their structure. Note that structural bistability has been ignored and only global minimum energies considered, since only five percent of octamers are bistable (Gardiner et al., 2003). A large number of octamer pairs that have identical or near identical minimum energy structures have very different structural tendencies (particularly with respect to roll). Pattern recognition via the structural probabilities may therefore find structural DNA fingerprints that would otherwise be unrecognised.

Chapter 5:
Structural Profiles – Single Sequence Queries

Structural profiles are graphical illustrations of how DNA structure varies across a sequence or set of sequences. They use the contents of the Octamer Database either to observe any characteristics of a single sequence that are special (a single sequence query) or to visualise a pattern common to a set of sequences (a multiple sequence query). They are an aid in understanding structural reasons for functional DNA activity, helping to answer questions, such as what structural features make a sequence have such a high affinity for a drug molecule or why does a protein recognise a particular set of DNA sequences? This chapter introduces the profiles that answer single sequence queries and presents Profile Manager (a software application developed to automate profile generation). For discussion of multiple sequence queries see Chapter 6.

A single sequence query is answered by a set of single sequence profiles. Each profile gives a graphical illustration of how a particular parameter varies across the sequence length with any special regions highlighted. Single sequence profiles can also be used to observe any striking similarities or differences between a sequence pair. Before presenting the profiles, a survey of the literature is made to identify any analogous tools that already exist.

5.1. Survey of Analogous Visualisation Tools

Attempts have been made to capture the nucleotide content of a DNA sequence graphically. One such example is a path followed in two-dimensional space, where each C, T, A and G refer to a movement north, south, west and east respectively (Randic and Vracko, 2000a). These graphical walks have been extended to three dimensions by describing a sequence of bases by movements along the vertices of a tetrahedron (Randic et al., 2000b). A Z-curve representation (Zhang et al., 2003) also exists, where a curve of N points (x_N, y_N, z_N) represents a sequence of length N (Equations 5.1 a, b and c). x_N represents the ongoing ratio of purine to pyrimidine bases, y_N represents the amino to keto ratio and z_N represents weaker hydrogen bond bases to stronger.

$$x_N = (A_N + G_N) - (C_N + T_N) \quad \textbf{Equ. 5.1a}$$

$$y_N = (A_N + C_N) - (G_N + T_N) \quad \textbf{Equ. 5.1b}$$

$$z_N = (A_N + T_N) - (G_N + C_N) \quad \textbf{Equ. 5.1c}$$

where A_N, C_N, G_N and T_N are cumulative counts of the number of respective bases encountered so far along the sequence.

Structural parameter plots of DNA can be obtained from the plot.it server (Vlahovicek et al., 2003) or from DNAssist (Patterton and Graves, 2000). Plot.it has 45 parameters to choose from (including roll, twist and tilt dinucleotide parameters and flexibility measures). Smoothing options are also available. The structural profiles presented here are more realistic than those of plot.it and DNAssist, since octamer units rather than context independent dinucleotide units are considered. Extra functionality in Profile Manager includes determination of unique areas of a sequence with respect to each parameter, the ability to summarise a set of profiles and an option to study the dynamics of DNA structure.

5.2. Single Sequence Profiles

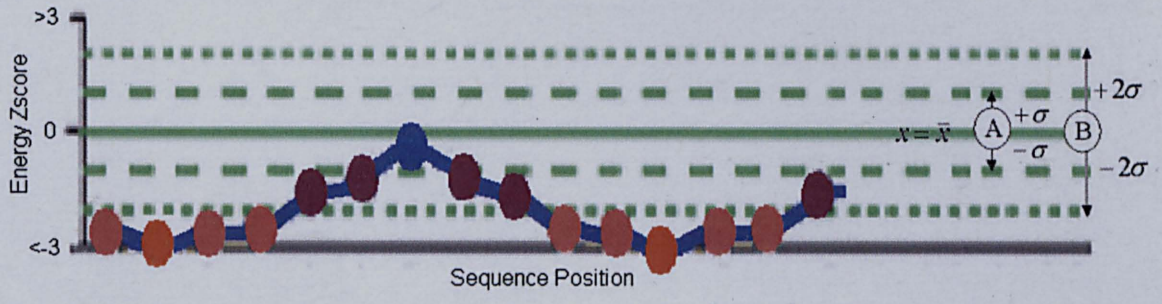
Consider a single DNA sequence and a single structural parameter. The first step in generating a structural profile is to convert the nucleotide sequence of length N into its consecutive overlapping (N-7) octamer sequence. For example the 10-letter sequence AACTTTGGTC is converted into 3 octamers: AACTTTGG, ACTTTGGT and CTTTGGTC. The chosen parameter's values are then retrieved from the octamer database for these octamer units. They are then each converted into a Zscore value that measures the importance/significance of a particular value of a parameter (Equation 5.2).

$$Zscore = \frac{x - \bar{x}}{\sigma} \quad \textbf{Equ. 5.2}$$

where x is a particular value under consideration, $\bar{x}$ is the mean of the parameter across the population of all possible octamers and σ is the population standard deviation.

A profile is then constructed by plotting the Zscore values against the sequence length (Figure 5.1). Cut-offs of <-3 and >3 at the minimum and maximum of the Zscore scale are used for visual purposes when comparing several profiles. Any parameters that fall outside this range are assigned values of –3 or +3 accordingly. The Zscore is the number of standard deviations a value is from the parameter's population mean. When considering a normal distribution, there is a 68% chance that a value will fall within plus or minus one standard deviation of the mean (region A in Figure 5.1) and a 95% chance that a value will fall within plus or minus two standard deviations (region B in Figure 5.1). Therefore any value that falls outside of these two boundaries (marked by green lines on a profile) is significantly different from average. Each value along a sequence has been colour coded, in order to highlight any special regions. At the two extremes, blue means average (within region A) and red means special (outside region B). Intermediate values, those between one and two standard deviations from the mean, are shown in purple.

Figure 5.1: *Example of a structural profile for a sequence's energy. Zscore boundaries are marked by green lines. For a normal distribution 68% of the data falls in region A and 95% in region B.*



The use of structural profiles to analyse the minimum energy structure and the flexibility of a sequence, in order to identify any interesting characteristics, is illustrated with two examples: the A-tract phenomenon and a Drosophila promoter comparison.

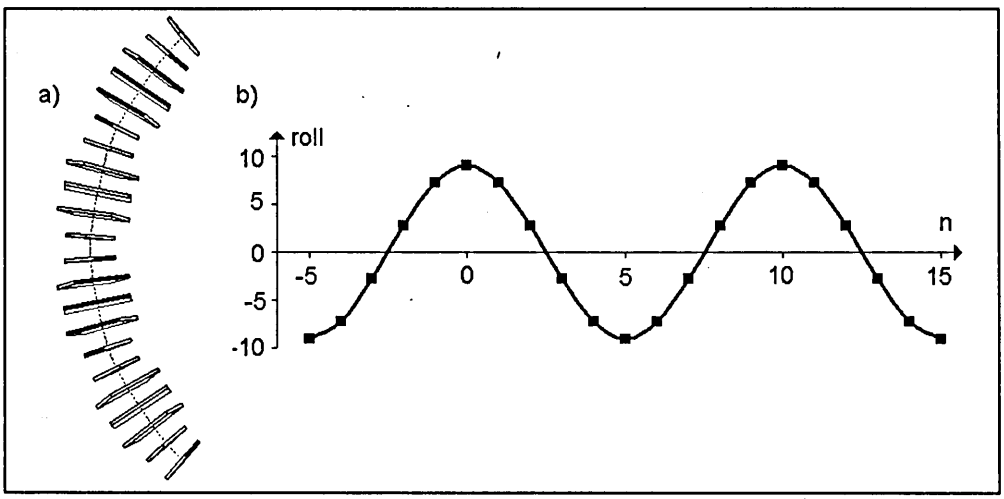
5.2.1. The A-tract Phenomenon

This example uses single sequence profiles to identify why two sequences that appear similar by their nucleotide composition are so different structurally (the A-tract phenomenon). An A-tract sequence is one that contains four or more adjacent adenine bases without a T-A step. The A-tract phenomenon refers to the difference between the

bent A-tract structure d(A_4T_4) and the straight structure d(T_4A_4). Note that d(S) means a sequence composed of repeating units of the subsequence S. What is it that makes a sequence curve? A sequence (Figure 5.2a) can accomplish a curvature of 45° per helical turn by using the periodic roll pattern shown in Figure 5.2b (Calladine and Drew, 2002). The roll at step n (R_n) varies as a cosine wave along the sequence (Equation 5.3).

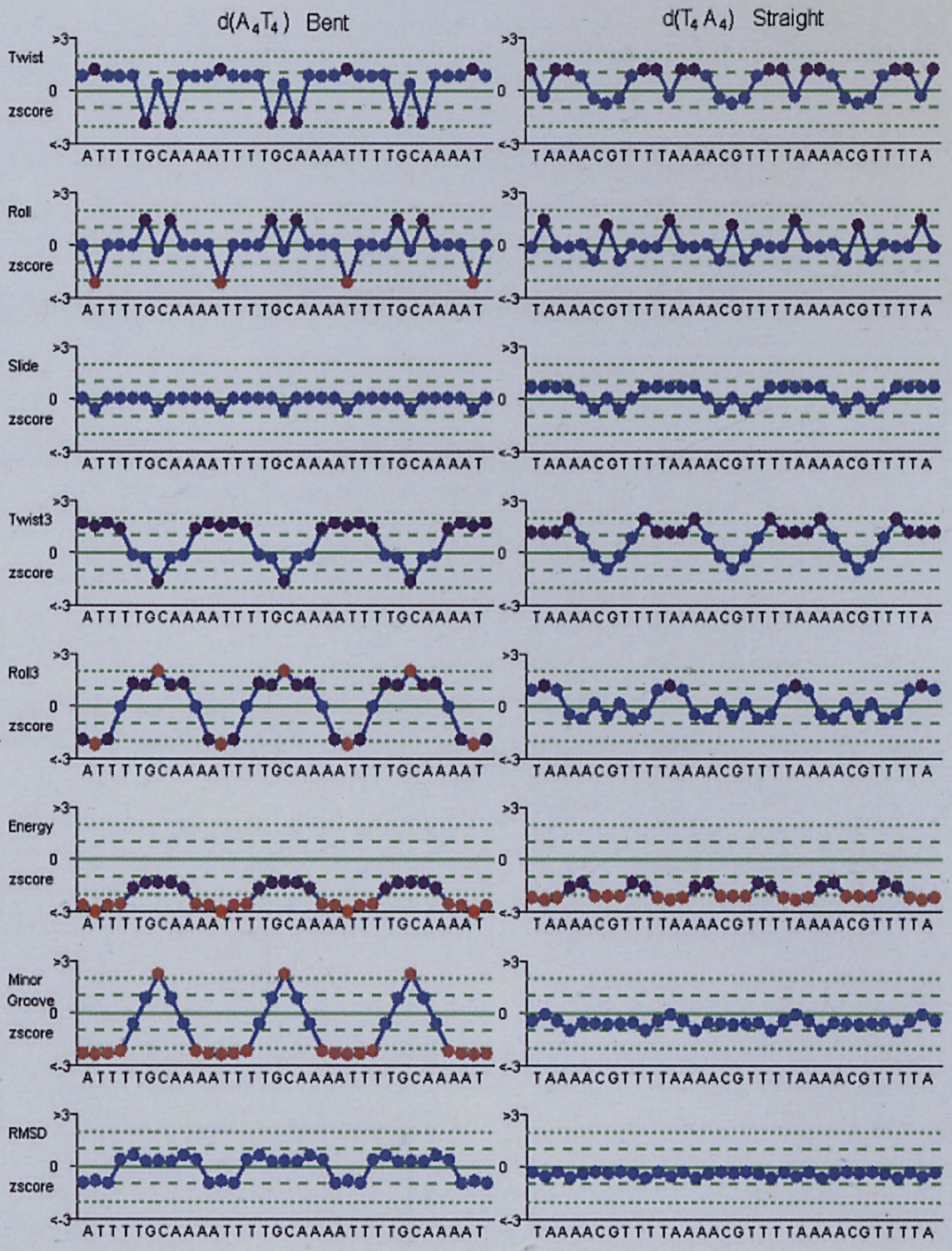
$$R_n = 9^\circ \cos(36^\circ n) \qquad \textbf{Equ. 5.3}$$

Figure 5.2: *Sequence curvature (Calladine and Drew, 2002) a) Sequence with 45° curvature per helical turn. b) A plot of the sequence's roll angle versus step number.*



Nuclear magnetic resonance structures of d(CA_4T_4G) and d(GT_4A_4C) have identified some interesting structural characteristics (Stefl et al., 2004), the majority of which are clearly illustrated with the structural profiles shown in Figure 5.3. Important features can be seen easily at a glance by focusing solely upon the shades of red, apparent in the roll, roll3, energy and minor groove profiles. The A-T steps of the d(A_4T_4) structure have large negative rolls, whereas the T-A steps of d(T_4A_4) have positive rolls in both the 1-step and 3-step roll profiles. The 10 base-pair periodic transitions between low and high 3-step roll in d(A_4T_4), reflect the smooth wave-like pattern of roll angles that cause DNA curvature (Calladine and Drew, 2002). Note that the energy of both sequences is extremely low across the majority of their lengths, meaning that these sequences are relatively stable. Narrow minor groove stretches interrupted by wide grooves at each of the G-C steps are found for d(A_4T_4). In comparison, nothing is special about the groove widths along d(T_4A_4).

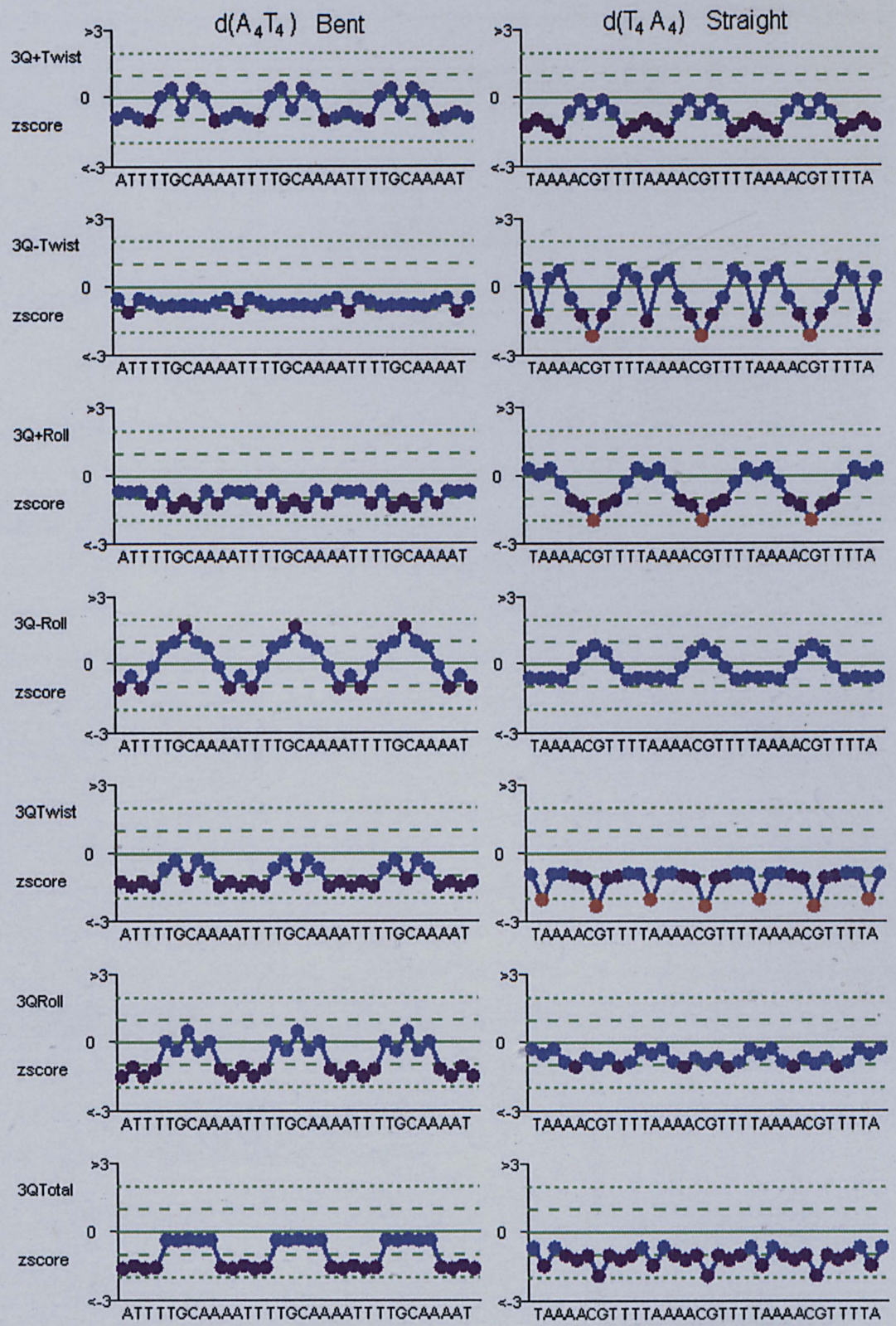
Figure 5.3: *Structure Profiles of the bent A-tract sequence d(A4T4) and the straight d(T4A4) sequence.*



Flexibility profiles in terms of the 3-step partition coefficients show no significantly flexible regions in either sequence (Figure 5.4). Significantly rigid steps can however be found in the d(T_4A_4) sequence. The C-G step is rigid in the decreasing twist, increasing roll and overall twist directions and the T-A step is rigid with respect

to overall twist. The d(A_4T_4) and d(T_4A_4) increasing twist flexibility profiles are almost identical. Looking at the $3Q_{Total}$ profiles it can be concluded that the overall flexibilities are similar.

Figure 5.4: *Flexibility Profiles of the bent A-tract sequence d(A_4T_4) and the straight d(T_4A_4) sequence.*



5.2.2. Drosophila Promoter Comparison

The Drosophila Core Promoter Database (Kutach and Kadonaga, 2000), publicly available online at www-biology.ucsd.edu/labs/Kadonaga/DCPD.html, contains 205 Drosophila Melangaster (fruit fly) promoters that are aligned by their experimentally determined transcription start sites. The three common promoter elements (discussed in Chapter 2) - the TATA-box (TATA), initiator (Inr) and downstream promoter element (DPE) - were identified along the promoters (Kutach and Kadonaga, 2000) and used to categorise them. The profiles of Figures 5.5 and 5.6 make an interesting comparison between two promoters from the database: ald (Shaw-Lee et al., 1992) and 4f-rnp (Petschek et al., 1997). Ald belongs to the DPE and TATA containing class and 4f-rnp to the class that possessed neither of these elements.

Two important positions (15 and 70) along both the promoters can be seen when focusing upon the common red areas. Both ald and 4f-rnp have a low slide, low 3-step twist, high RMSD and fluctuations in 3-step roll at positions 15 and 70 (Figure 5.5). High energy is also a common feature at 70. These patterns show clear agreement between the structural alignment of the promoters and alignment by their experimental transcription start sites, suggesting that certain structural features could be used for promoter recognition, even across the different classes of promoters. Multiple sequences should however be considered before making any solid conclusions, see Chapter 6.

A large transition of high to low decreasing twist flexibility ($3Q^{-}_{twist}$) is present in both promoters between positions 65 and 70 (Figure 5.6). Octamers that are very flexible by increasing roll appear at the ends of the sequences. Additional large transitions are present, but not in common positions. Sudden changes in flexibility may therefore be an important promoter feature. Flexible octamers may put stress upon the surrounding rigid octamers and present sites along a sequence where the double helix can be easily unravelled for transcription initiation. This hypothesis among others will be further investigated in Chapter 6.

Figure 5.5: *Structural Profiles of two Drosophila promoters. TATA=TATA-box, Inr=initiator and DPE=downstream promoter element. The experimentally determined transcription start site is at position 47. (a) The Ald promoter. (b) The 4f-rnp promoter.*

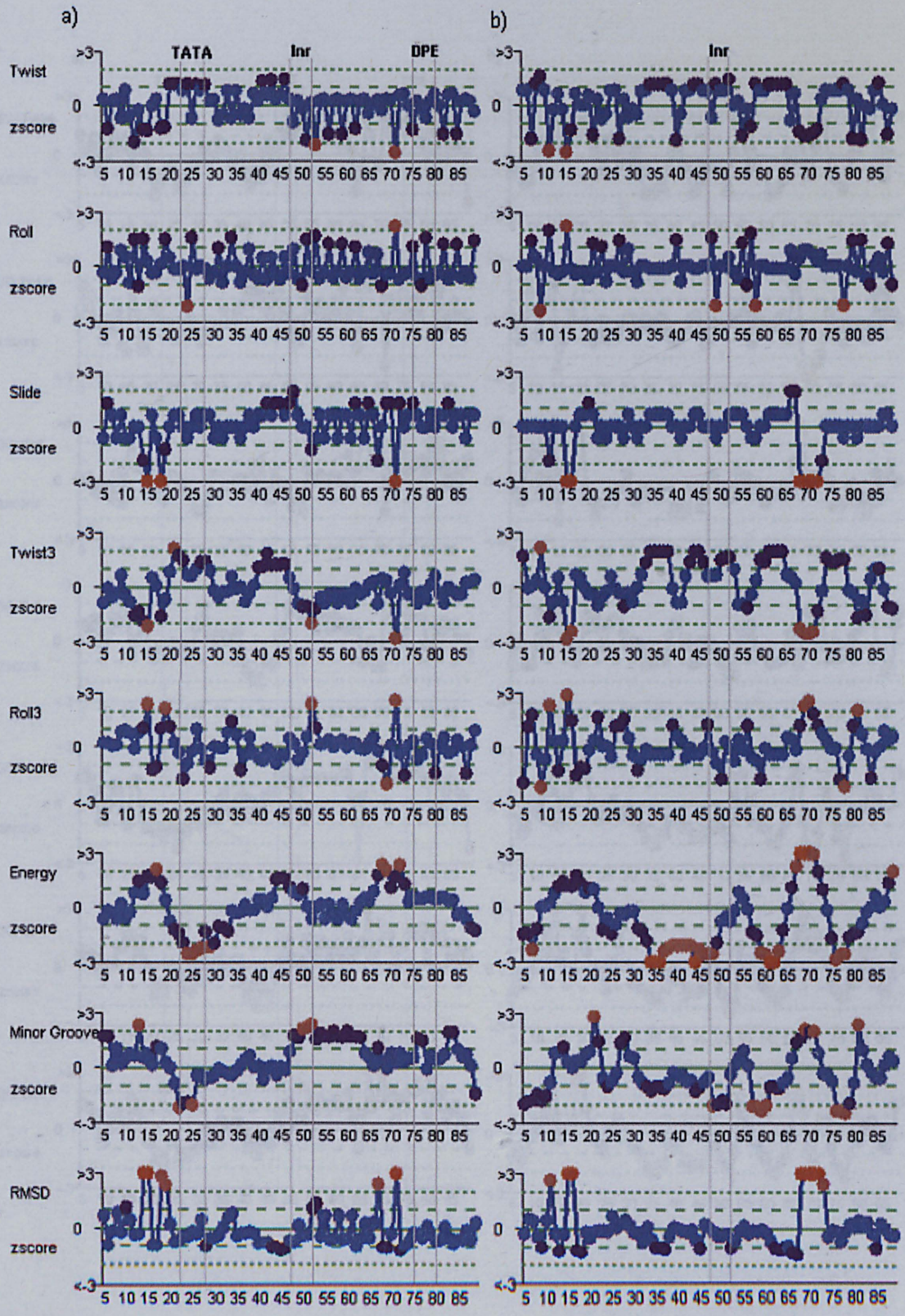
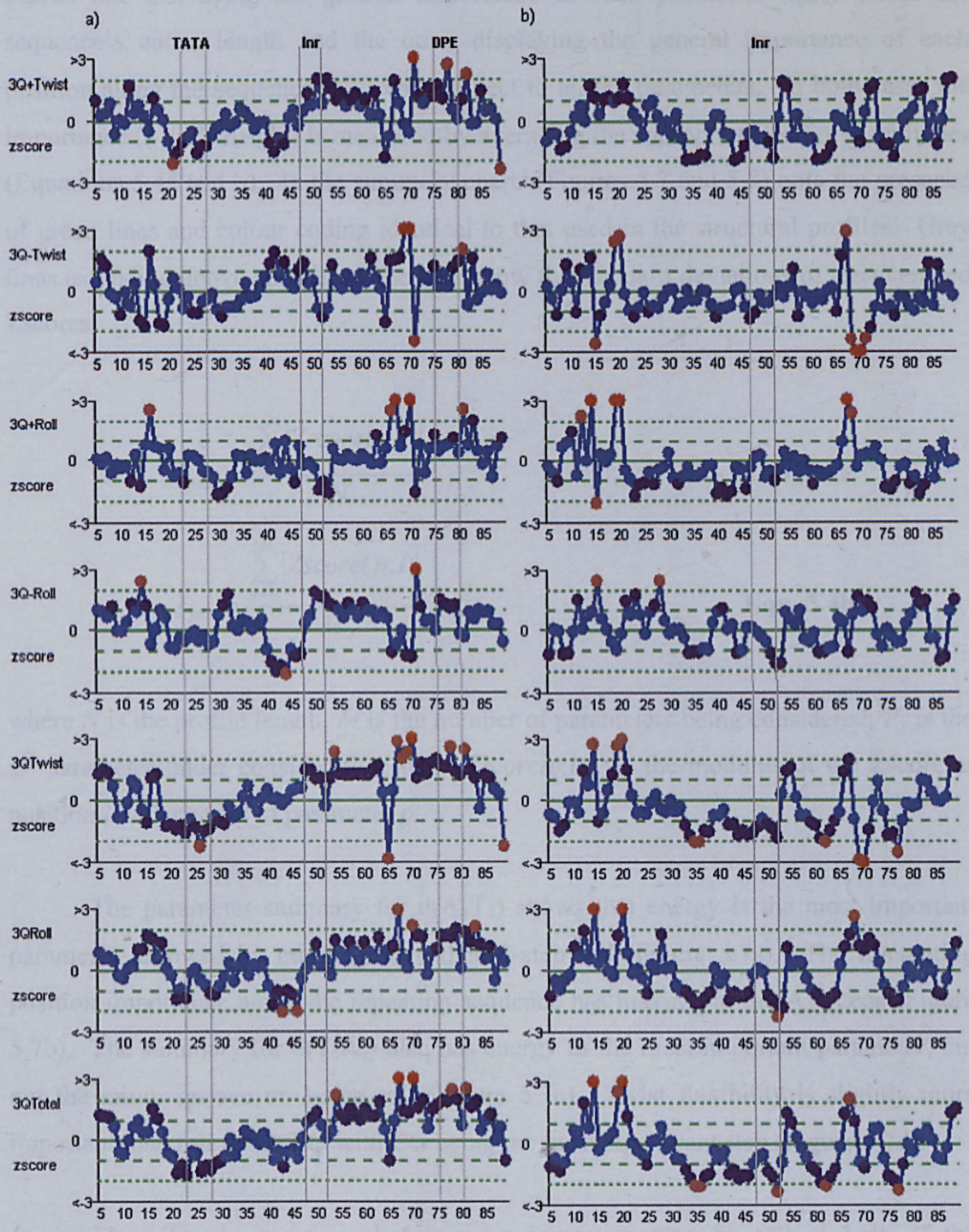


Figure 5.6: *Flexibility Profiles of two Drosophila promoters. TATA=TATA-box, Inr=initiator and DPE=downstream promoter element. The experimentally determined transcription start site is at position 47. (a) The Ald promoter. (b) The 4f-rnp promoter.*



5.3. Summary Charts

The information contained within a set of profiles can be summarised by two bar charts: one displaying the general importance of each parameter (I_{para}) across the sequence's entire length and the other displaying the general importance of each position along the sequence (I_{pos}) with respect to all the parameters. In both cases the importance (a bar's height) is measured by averaging the appropriate Zscore magnitudes (Equations 5.4a and b). In the summary charts (Figures 5.7 and 5.8) note the presence of green lines and colour coding identical to that used in the structural profiles. Grey lines extending above and below the bars show the standard deviations of the averaged Zscores.

$$I_{para} = \frac{\sum_{i=1}^{N} |Zscore(p,i)|}{N} \qquad \textbf{Equ. 5.4a}$$

$$I_{pos} = \frac{\sum_{p=P_1}^{P_M} |Zscore(p,i)|}{M} \qquad \textbf{Equ. 5.4b}$$

where N is the profile length, M is the number of parameters being considered, P_x is the x^{th} parameter under consideration and $|Zscore(p,i)|$ is the modulus of the Zscore at position i with respect to parameter p.

The parameter summary for $d(A_4T_4)$ shows that energy is the most important parameter followed by minor groove then 3-step roll (Figure 5.7a). The fluctuating position importance along the repeating sequence has maxima at the AT steps (Figure 5.7b). The summary for $d(T_4A_4)$ also has energy as the most important parameter, but not the minor groove or 3-step roll (Figure 5.7c). Twist flexibility is slightly more important than roll flexibility with CG being the most significant step (Figure 5.7d).

Flexibility by twist is slightly more important than by roll in both of the promoter sequences (Figure 5.8 a and c). Energy and 3-step twist are generally more important in 4f-rnp than ald. Clear peaks are seen in the position summaries at around 15 and 70, indicating areas where common features may be present (Figure 5.8 b and d).

Figure 5.7: *A-tract phenomenon summaries. a) d(A_4T_4) parameter summary, b) d(A_4T_4) position summary, c) d(T_4A_4) parameter summary and d) d(T_4A_4) position summary*

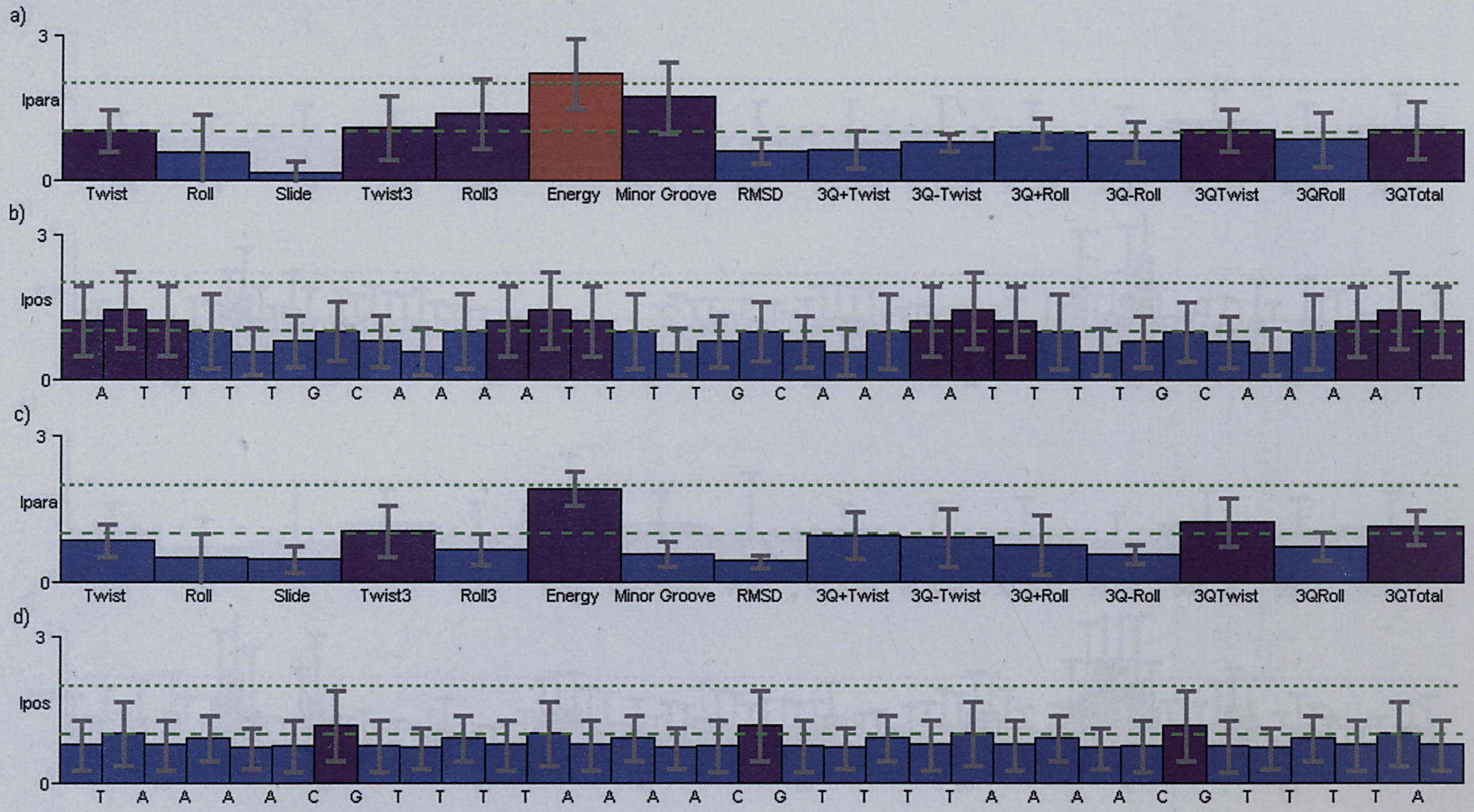
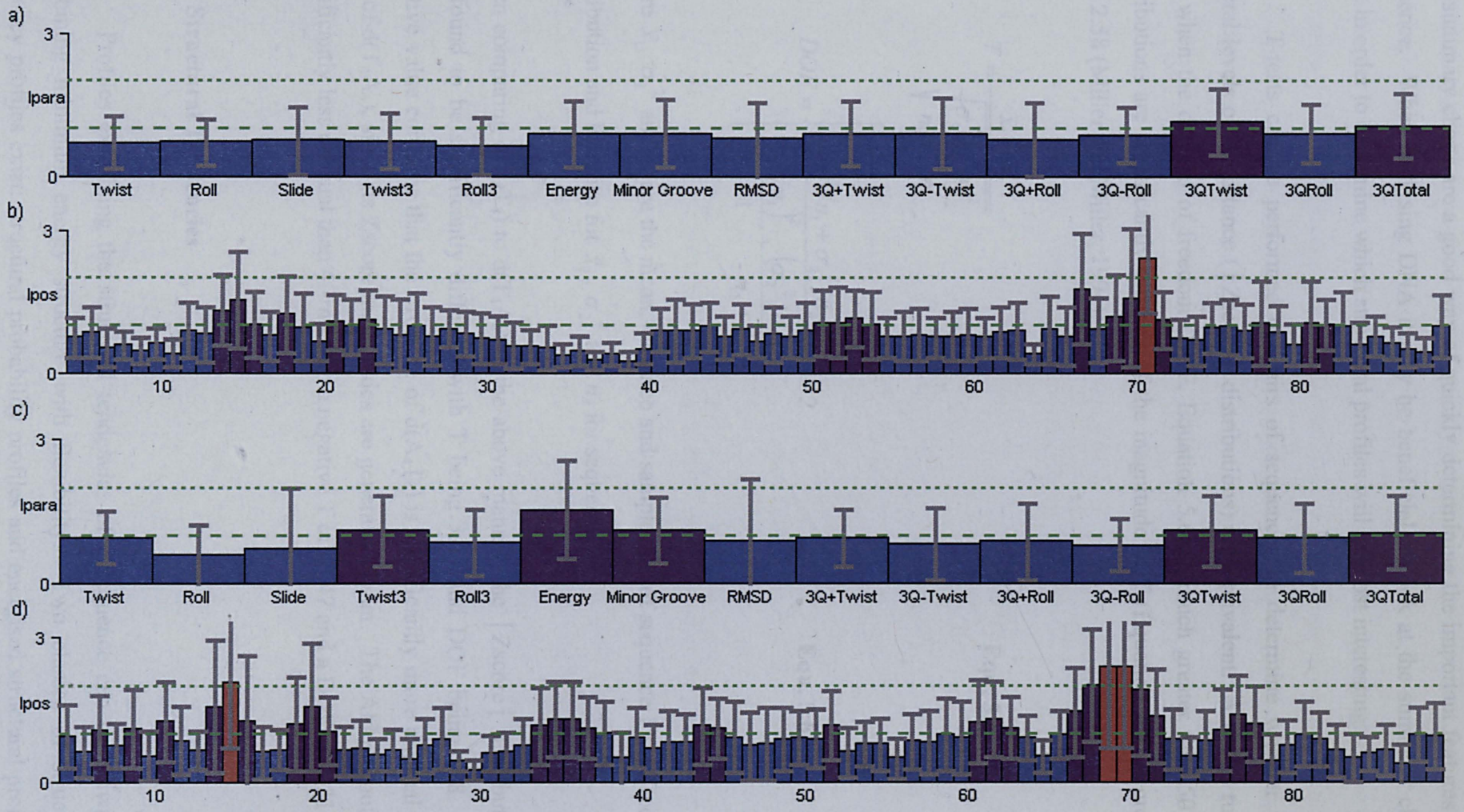


Figure 5.8: *Promoter sequences summaries. a) Ald parameter summary, b) Ald position summary, c) 4f-rnp parameter summary and d) 4f-rnp position summary*



The summary charts are a good way of quickly determining the important features of a sequence. When analysing DNA it may be beneficial to look at the summary charts first, in order to determine which structural profiles will be most interesting.

T-tests can be performed on pairs of sequences to determine whether their general levels of importance ($|\text{Zscore}|$ distributions) are equivalent. A useful rule is that when the degrees of freedom (DOF, Equation 5.6) is much greater than 50, the distributions are significantly different if the magnitude of T (Equation 5.5) is greater than 2.58 (Miller and Miller, 1994).

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}} \qquad \textbf{Equ. 5.5}$$

$$DOF = \left\{ \frac{\left({\sigma_1}^2/n_1 + {\sigma_2}^2/n_2\right)^2}{\frac{\left({\sigma_1}^2/n_1\right)^2}{n_1+1} + \frac{\left({\sigma_2}^2/n_2\right)^2}{n_2+1}} \right\} - 2 \qquad \textbf{Equ. 5.6}$$

where $\bar{x}_1$, ${\sigma_1}^2$ and n_1 are the mean, variance and sample size of sequence 1's $|\text{Zscore}|$ distribution and likewise for $\bar{x}_2$, ${\sigma_2}^2$ and n_2 for sequence 2.

When comparing d(A_4T_4) to d(T_4A_4) in the above manner, the $|\text{Zscore}|$ distributions are found to be significantly different with T being 3.54 and DOF being 928. T's positive value confirms that the structure of d(A_4T_4) is significantly more unusual than that of d(T_4A_4), since its Zscore magnitudes are generally higher. The Ald promoter is significantly less unusual than 4f-rnp with a negative T of –3.47 and a DOF of 2431.

5.4. Structural Tendencies

Profiles examining the structural tendencies of a sequence can be drawn by combining minimum energy structure with flexibility. Two classes of structural tendency profiles exist: structural probability profiles and energised structural profiles.

Structural probability profiles approximate the range of roll or twist around the minimum energy structure that an octamer populates by a specified probability. Calculations are based on the theory of Chapter 4, estimating the probability that an octamer has a roll between a and b, $P[a \le r \le b]$. The upper and lower structural boundaries for a particular octamer are determined by starting at the energy minimum value and making one degree adjustments to the boundaries until the desired probability (P_{CUTOFF}) has been reached. The pseudocode is given below:

1. Set P_{CUTOFF} to desired value (default is 0.75)
2. Set lower and higher to the minimum energy roll
3. If $P[lower - 1 \le r \le higher] > P[lower \le r \le higher + 1]$

   then lower = lower –1

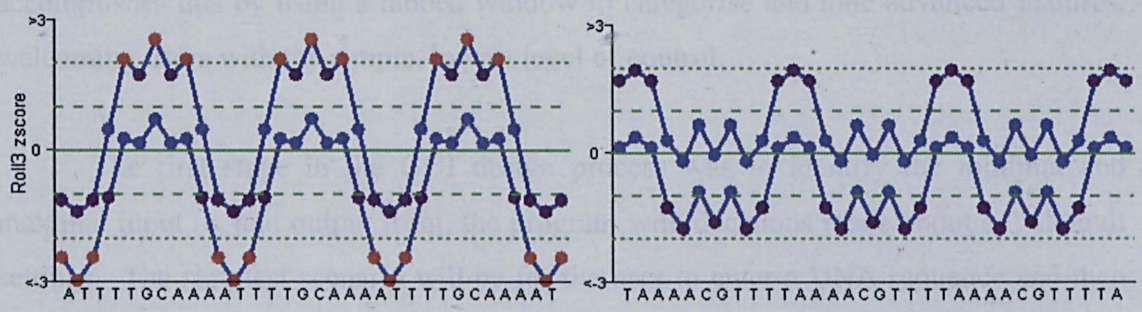
   Else if $P[lower - 1 \le r \le higher] < P[lower \le r \le higher + 1]$

   then higher = higher +1

   Else lower = lower – ½ and higher = higher + ½
4. If $P[lower \le r \le higher] \ge P_{CUTOFF}$ then STOP else RETURN to step 3

A structural tendencies profile is drawn with a lower and upper curve, representing the structural boundaries. The space between the curves is the populated structural space. The 3-step roll structural tendencies have been examined for the $d(A_4T_4)$ and $d(T_4A_4)$ sequences (Figure 5.9). A periodic roll curve can be formed for both sequences, although with a much smaller magnitude for $d(T_4A_4)$ than $d(A_4T_4)$. This suggests that although $d(T_4A_4)$ has a straight minimum energy conformation it can be bent to a certain, but much lesser, extent than the already highly curved $d(A_4T_4)$.

Figure 5.9: *Structural 3-step roll tendencies of $d(A_4T_4)$, on the left, and $d(T_4A_4)$, on the right, using a probability cut-off of 0.75*



Energised structural profiles are an alternative way of viewing structural tendencies. A chosen amount of energy is applied to each octamer in both directions from the energy minimum structure. Lower and upper structural boundaries can then be plotted as before, and are determined using the force constants (Equations 5.7a and b).

$$lower = r_{\min} - \sqrt{E / k^{-}_{roll}} \qquad \textbf{Equ. 5.7a}$$

$$upper = r_{\min} + \sqrt{E / k^{+}_{roll}} \qquad \textbf{Equ. 5.7b}$$

5.5. Profile Manager

Profile Manager is an application currently under development that aims to automate profile generation in an efficient, user-friendly environment. The design of Profile Manager version 1 (v.1) is explained with details about the graphical user interface (GUI).

The GUI is the communication device between the end-user of the application and the computer. It therefore needs to be both easy to understand and fully functional. These two factors may compromise one another, since the more functionality a GUI offers, the more complicated it may appear, and the more expert user knowledge it will require. Mandel makes an analogy of user control to the choice between taking a train and driving a car (Mandel, 1997). The car driver refers to the expert who desires full control. The train passenger is the novice who wants to be taken to the answer with minimal knowledge of how they got there. Any good application should be designed for both novices and experts. Therefore Profile Manager tries to give experts the option to "drive" without baffling the novice with hundreds of controls. Profile Manager accomplishes this by using a tabbed window to categorise and hide advanced features, welcoming users with the simple, lowest level of control.

The first stage in the GUI design process was to identify the minimal and maximal input to, and output from, the program with decisions made about any default settings. The simplest scenario will be for the user to enter a DNA sequence and then press a button and receive a set of profiles. More complicated scenarios can be

understood by identifying variables on which the output depends. Table 5.1 lists the attributes associated with a collection of profiles and Table 5.2 gives details of any additional functionality that the GUI interface should have (such as a way of exiting the application).

Table 5.1: *Variables associated with a collection of profiles*

Variable	Comment
List of parameters	Dependent upon profile type: • Minimum energy structure profile *Central 1-step & 3-step parameters and ground state properties* • Flexibility profile *The partition coefficients* • Structural probability profile *1-step & 3-step twist and roll* • Energised structural profile *1-step & 3-step twist and roll*
Sequence	Entered directly or via file
Background distribution	Parameter population distributions as the default or based on sequence/sequences given in a file.
Special sequence positions	To be highlighted in plot, i.e. transcription start site
Maximum sequence frame length	Maximum sequence length covered by a set of profiles. If length exceeds this then continue in a second window of profiles
x-axis labels	Label by the nucleotides or nucleotide position
y-axis labels	Label by the z-score values or parameter values
y-axis scale	Define how many z-scores to scale by or use parameter ranges

Table 5.2: *Additional functions of Profile Manager v.1*

Command name	Function
Refresh	To reset variables to their default settings
Exit	To exit application
Help	To aid users
Include Summary	To draw a bar chart to summarise profiles
Draw Profiles	To draw the selected profiles

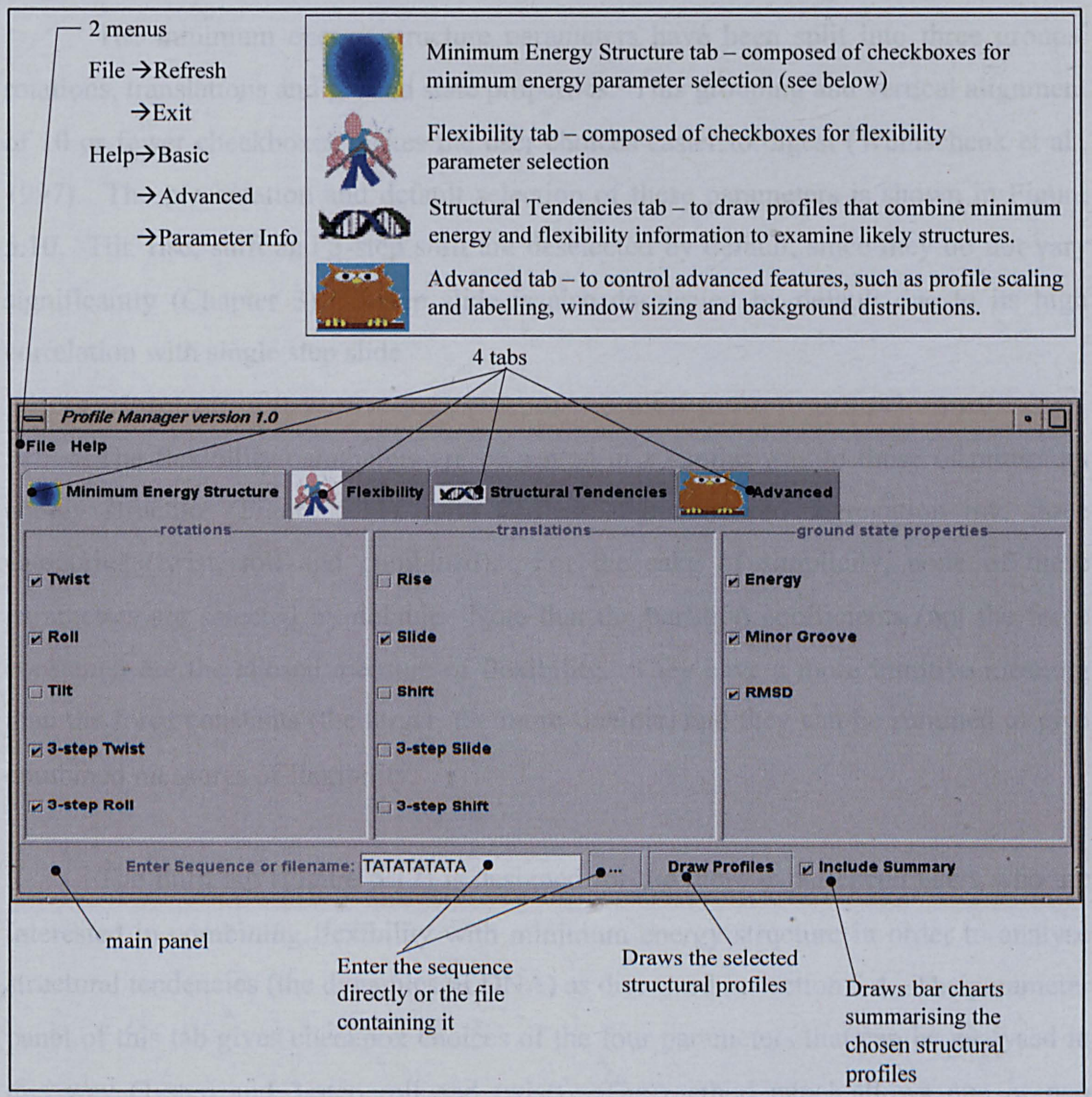
Each of the variables has a default value and a control for its manipulation. The location and type of a control are important factors to consider and affect the GUI interface design. The commands of Table 5.2 also need controls to activate their functions. A summary of the control type and its location for each variable and command is tabulated in Table 5.3. This analysis resulted in a GUI composed of two menus and four tabs positioned on a main panel (Figure 5.10).

Table 5.3: *Location and type of all the GUI components*

Variable / Command	Control Type	Control Location
Parameter	Checkbox	Tab panel categorised by profile type
Sequence	Text box and [...] button	Main panel
Background distribution	Text box and [...] button	Advanced tab
Special sequence positions	Text box	Advanced tab
Maximum sequence frame length	Slider	Advanced tab
x-axis labels	Pair of radio buttons	Advanced tab
y-axis labels	Pair of radio buttons	Advanced tab
y-axis scale	Slider & pair of radio buttons	Advanced tab
Refresh	Menu item	File menu
Exit	Menu item	File menu
Help	Menu item	Help menu
Include summary	Checkbox	Main panel
Draw Profiles	Button	Main panel

The four tabs refer to three profile categories (minimum energy structure, flexibility and structural tendencies) plus an advanced tab. Note that the structural probability profiles and energised structural profiles have been grouped together as structural tendencies, since they both use the same parameters to describe likely structures. Separation of the parameters across a tabbed panel is essential, as it would be overwhelming to present them together and a user (particularly a novice) may only be interested in one profile type at a given time. The most common and simple profile type will be that of minimum energy structure; therefore, this is the default tab displayed when the application is opened (Figure 5.10).

Figure 5.10: *Profile Manager's GUI and the Minimum Energy Structure tab*



The menu bar is located above the tabbed panel and is composed of a File and Help menu. Refresh and Exit are tucked away within the File menu, since although they are common actions they are also destructive and do not want to be activated by mistake with a single click of a button. The Help menu allows easy selection of the different types of available help (basic, advanced and parameter information). Components belonging to the main panel represent frequent actions that can be seen regardless of the current active tab. The sequence or the name of the file containing the sequence can be entered into the text box. Alternatively the [...] button can be used to select the file from directory listings. The [Draw Profiles] button gets the profiles with optional summary bar charts specified by the Include Summary checkbox.

The minimum energy structure parameters have been split into three groups: rotations, translations and ground state properties. This grouping and vertical alignment of 10 or fewer checkboxes makes the user choices easier to digest (Weinschenk et al., 1997). The presentation and default selection of these parameters is shown in Figure 5.10. Tilt, rise, shift and 3-step shift are deselected by default, since they do not vary significantly (Chapter 3). 3-step slide is also deselected by default due to its high correlation with single step slide.

The flexibility parameters are presented in a similar way to those of minimum energy structure (Figure 5.11), with vertical alignment and segregation into three categories (twist, roll and combined). For the sake of simplicity, none of these parameters are selected by default. Note that the partition coefficients (not the force constants) are the chosen measure of flexibility. They have a more intuitive meaning than the force constants (the larger, the more flexible) and they can be summed to give combined measures of flexibility.

The third tab (Figure 5.12) is designed for the more experienced users who are interested in combining flexibility with minimum energy structure in order to analyse structural tendencies (the dynamics of DNA) as discussed in section 5.4. The parameter panel of this tab gives checkbox choices of the four parameters that can be analysed in this way (1-step and 3-step roll and twist). The method panel allows one of two methods to be selected: probabilistic to obtain structural probability profiles or energetic to obtain energised structural profiles. The third panel titled "cut-off values" gives

slider controls to select either the probability cut-off or energy cut-off, depending on the selected method. Note that the appropriate slider is greyed out, so as to make it clear to the user that this control has no use for the current selected method. The default probability cut-off is 0.75 and the default energy cut-off is 1 $kJmol^{-1}$.

Figure 5.11: *Profile Manager's Flexibility tab*

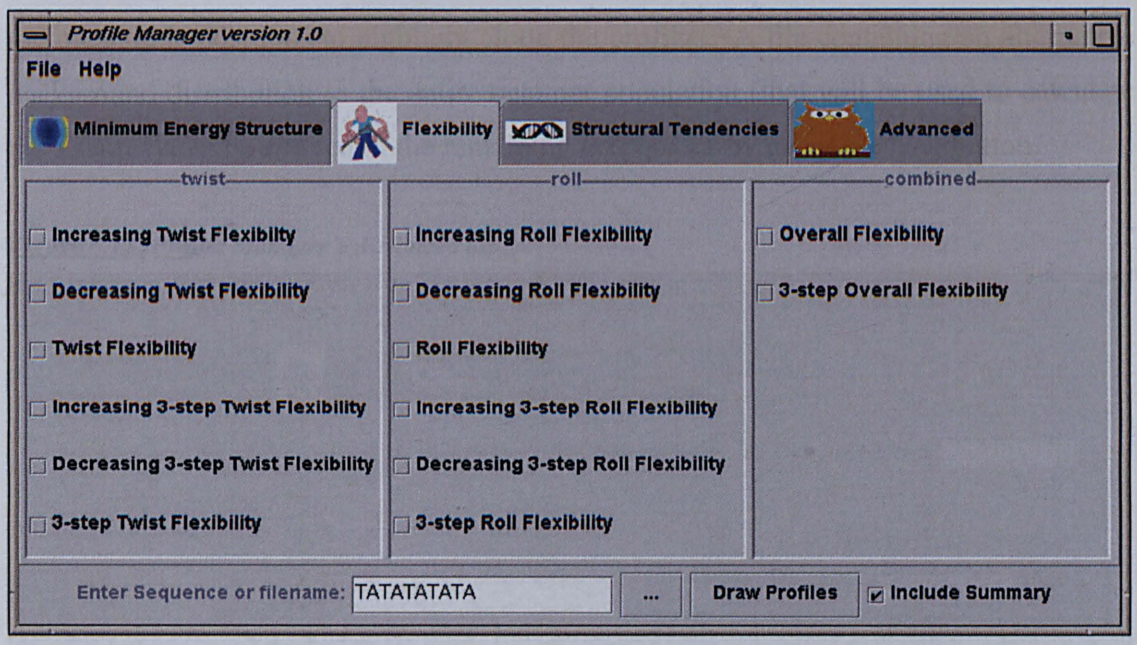
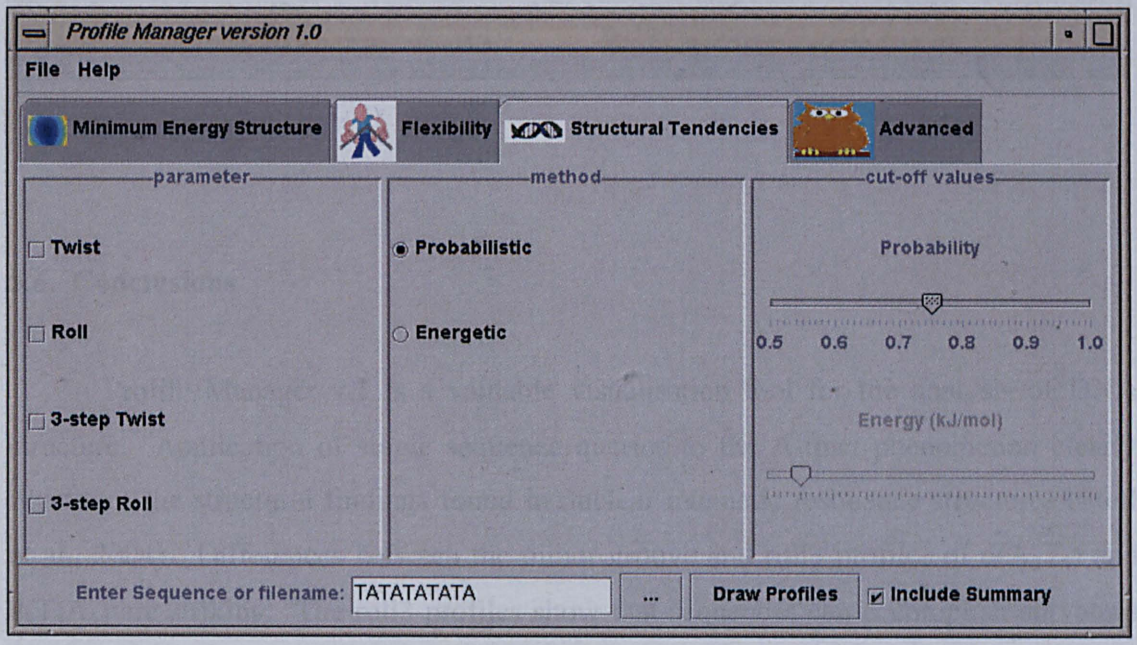
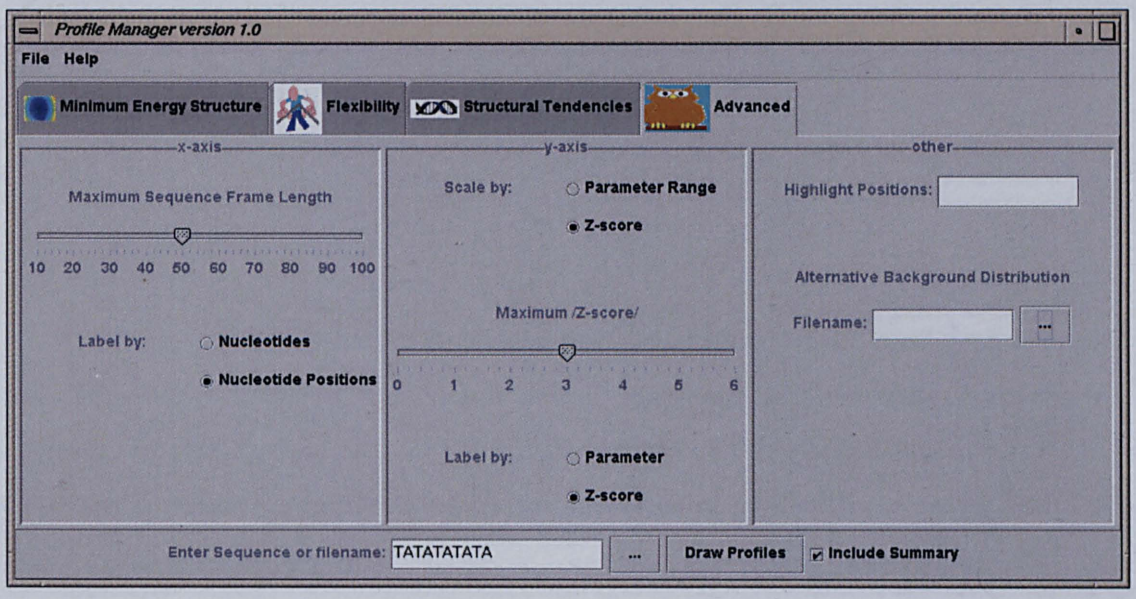


Figure 5.12: *Profile Manager's Structural Tendencies tab*



The Advanced tab (Figure 5.13) is split into an 'x-axis' panel, 'y-axis' panel and 'other' panel. The 'x-axis' panel contains a slider control for the maximum sequence frame length (whose default value is 50) and a radio button choice of nucleotides or nucleotide positions for the x-axis labels. The 'y-axis' panel contains radio buttons for the y-axis scale and label choices. A slider is also present for manipulation of the maximum |Zscore| value (default of three). The 'other' panel contains miscellaneous advanced options. The 'Highlight Positions' text box allows comma-separated input of the nucleotide positions to highlight along the profiles. A file containing an alternative background distribution to the entire octamer population (that will be used to calculate the Zscores) can be entered in the remaining text box or by using the [...] button.

Figure 5.13: *Profile Manager's Advanced tab*



5.6. Conclusions

Profile Manager v.1 is a valuable visualisation tool for the analysis of DNA structure. Application of single sequence queries to the A-tract phenomenon clearly illustrates the structural findings found in nuclear magnetic resonance structures (Stefl et al., 2004). Differences between the minor groove and roll3 profiles of $d(A_4T_4)$ and $d(T_4A_4)$ are striking. The roll3 profiles show that sequences can accomplish curvature by using a periodic roll pattern. Two important positions (15 and 70) and frequent

sudden changes in flexibility were identified in both of the studied promoters. Multiple sequences should now be studied to determine if similar patterns are common to promoters in general (Chapter 6).

The summary charts illustrate how the general importance of a sequence varies with respect to the parameters or how the importance varies across a sequence for a combination of parameters. Significant differences in the importance of two sequences can be assessed by a T-test. The structural tendency profiles provide an insight into structural dynamics. However, none of these approaches are capable of pattern recognition across multiple sequences, the subject of the remaining chapters.

Chapter 6:
Structural Profiles – Multiple Sequence Queries

The use of structural profiles to answer multiple sequence queries is explored. This work leads on from Chapter 5, where single sequence profiles were presented and used to answer single sequence queries. Initially, sequence logos are introduced. A logo is the equivalent visualisation tool to structural profiles that is used to summarise the nucleotide patterns in a set of sequences. The remainder of this Chapter uses structural profiles to explore patterns in flexibility for a set of promoters. Bendability profiles have previously been generated to analyse flexibility patterns in a set of pre-aligned promoter sequences (Pedersen et al., 1998). The analogous structural profiles are generated here with the roll and twist flexibility parameters.

6.1. Sequence Logos

First, consider a multiple sequence alignment that will be displayed throughout this research by a matrix plot. Each row of a matrix represents a sequence and each column represents a particular nucleotide position in the alignment. Therefore an element in the matrix represents a single nucleotide or gap, which is colour coded blue for C, orange for G, green for A, red for T and black for a gap. An example of a matrix plot is given in Figure 6.1a.

A sequence logo (Schneider and Stephens, 1990) takes each column in a multiple sequence alignment and turns it in to a frequency distribution of bases, which it then displays as a stack of letters (Figure 6.1b). The heights of the letters within a single stack are proportional to their relative frequency within a column (Equation 6.1). The letters are ordered so that the most frequent appears at the top of the pile and vice versa. Therefore a consensus sequence can be obtained by reading across the top letters in a logo plot. The total height of a letter stack (R_L, Equation 6.2) equals the importance of that nucleotide position relative to the other alignment columns. This importance is otherwise known as the column's information content and is measured in bits. It can be

viewed as a loss in uncertainty, hence the two entropy terms, S_{BEFORE} and S_{AFTER}. Gaps are not displayed in sequence logos, but they suppress the heights of corresponding letter stacks. A clear explanation of information content, uncertainty and bits is given by Shaner et al. (1993). Note that all sequence logos displayed within this research were obtained using the online tool WebLogo (Crooks et al., 2004) at http://weblogo.berkeley.edu.

$$h_{b,L} = f(b,L)R_L \qquad \textbf{Equ. 6.1}$$

where $h_{b,L}$ is the height of nucleotide b in column L, $f(b,L)$ is the frequency of base b in column L and R_L is the information content of the letter stack for column L.

$$R_L = S_{BEFORE} - S_{AFTER} \qquad \textbf{Equ. 6.2a}$$

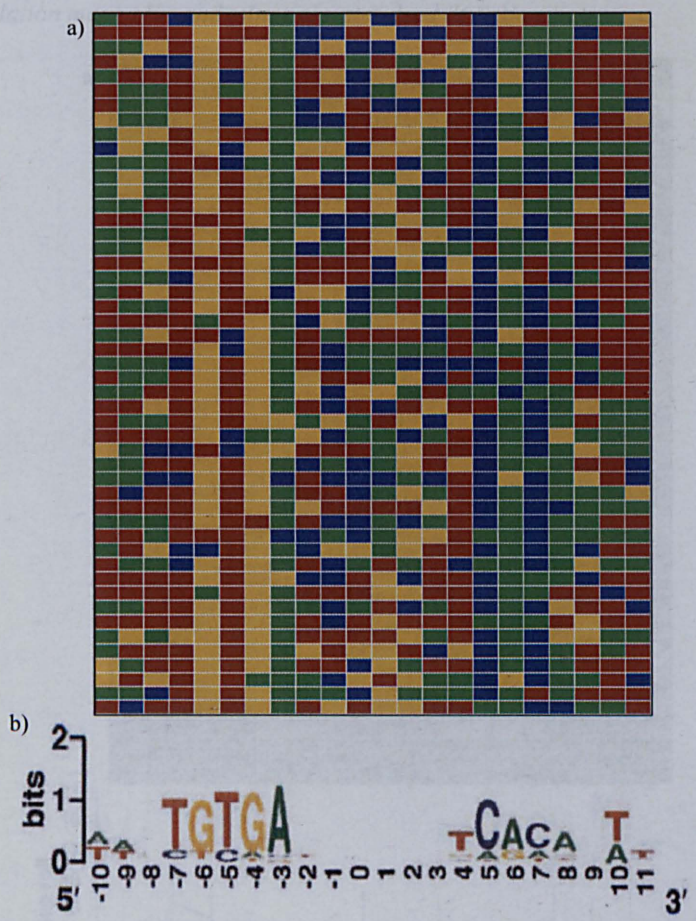
$$S_{BEFORE} = \log_2 N \qquad \textbf{Equ. 6.2b}$$

$$S_{AFTER} = -\sum_{b=1}^{N} f(b,L)\log_2 f(b,L) \qquad \textbf{Equ. 6.2c}$$

where S_{BEFORE} is the entropy before the alignment and S_{AFTER} the entropy after the alignment. N is the alphabet size (4 for DNA) and $f(b,L)$ is the frequency of base b in column L.

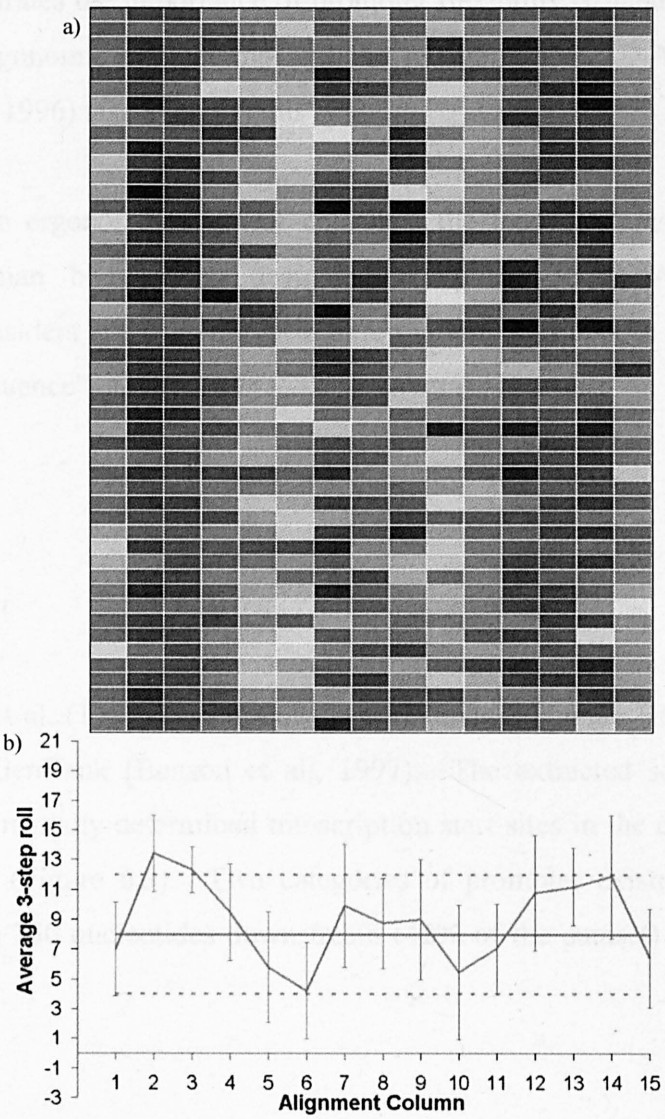
Nucleotide patterns within a set of Catobolite Activator Protein (CAP) binding sites were observed via a matrix plot (Figure 6.1a) and a sequence logo (Figure 6.1b). The double hump in the logo is associated with the fact that CAP is a homodimeric protein that has two helix-turn-helix motifs (Schultz et al., 1991). As mentioned in Chapter 2 section 2.2, CAP's two recognition helices bind to consecutive turns of the major groove, bending the DNA by 90°. Common patterns in the structure and how easily it can be bent are clearly important to these sequences. The consensus sequence (the nucleotides positioned at the top of the letter stacks) is a palindrome. Two large kinks in the structure have been identified as occurring at the TG/CA base pairs that are 5-6 positions on either side of the dyad axis (Schultz et al., 1991).

Figure 6.1: *Nucleotide patterns within a set of Catobolite Activator protein binding sites.*
a) The matrix plot, where each row is a sequence and each column is a position along the alignment. Blue = C, orange = G, green = A and red = T. b) Sequence logo, obtained from Weblogo (Crooks et al., 2004).



Assuming that the set of CAP binding sites are pre-aligned by their structure, a matrix plot illustrating their 3-step roll alignment patterns can be drawn (Figure 6.2a). Here, each element of the matrix represents an octamer's 3-step roll rather than a nucleotide and is shaded from white to black meaning low to high 3-step roll respectively. The structural matrix plot can be summarised by a structural profile (Figure 6.2b), showing the variation in the average 3-step roll along the sequences. Standard error bars are given at each alignment position and represent the variation of roll down a particular alignment column. The solid central green line represents the mean of 3-step roll across the entire octamer population and the dotted green lines show this mean plus or minus one population standard deviation. The roll pattern is clearly symmetric with the two kinks present at octamer positions 2 and 14.

Figure 6.2: *3-step roll patterns within a set of Catobolite Activator protein binding sites.*
a) The matrix plot, where each row is a sequence and each column is a position along the alignment. Light to dark shading meaning low to high 3-step roll. b) 3-step roll structural profile, showing the variation of average 3-step roll (measured in kJmol^{-1}degrees^{-2}) along the sequence. The green lines refer to the octamer population mean plus and minus one standard deviation.



6.2. Promoter Flexibility Case Study

Characteristic DNA flexibility patterns have previously been found in a set of pre-aligned promoter sequences via average bendability profiles (Pedersen et al., 1998). Three flexibility measures were used that all independently identified the same general trend. This case study explores the analogous roll and twist flexibility profiles.

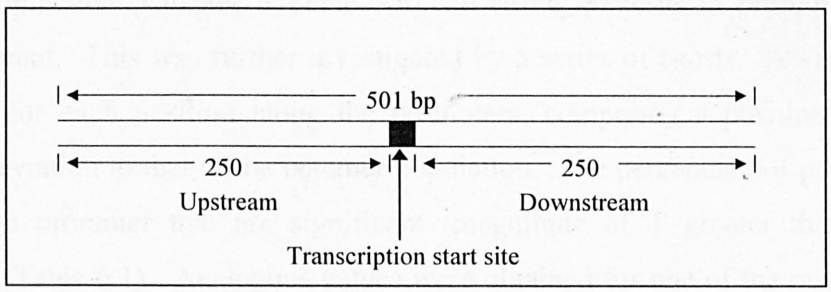
Features common to the dataset but not to random sequences will form a structural fingerprint of promoter activity. The analysis of patterns in flexibility along promoters is justified by the single promoter profiles of Chapter 5, where frequent transitions in flexibility were found along two promoters. The TBP-TATA complex is a classic example that illustrates the importance of promoter flexibility (Chapter 2). An analogy made between ergonomic engineering and the recognition of DNA by transcription factors (Juo et al., 1996) also supports this study.

> "An ergonomic engineer considers the local motions possible for the human body when designing equipment; a DNA-binding protein considers and uses the local deformations available to a particular target sequence".
>
> (Juo et al., 1996)

6.2.1. The Dataset

Pedersen et al. (1998) selected 624 non-redundant Human RNA Polymerase II promoters from GenBank (Benson et al., 1997). The extracted sequences are pre-aligned by experimentally determined transcription start sites in the centre of their 501 base pair lengths (Figure 6.3). Two categories of promoter exist: those containing codons within the 250 nucleotides downstream (42% of the dataset) and those that do not.

Figure 6.3: *The Promoter Template. Each sequence has its experimentally determined transcription site located in the centre with 250 nucleotides upstream and downstream.*



Three flexibility models were used to generate the profiles: (1) a tri-nucleotide DNase I cutting frequency model (Brukner et al., 1995); (2) a tri-nucleotide model based on the location preferences of nucleosomes (Satchwell et al., 1986; Goodsell and Dickerson, 1994); and (3) a dinucleotide propeller-twist model, related to slide mobility as discussed in Chapter 3 (ElHassan and Calladine, 1996). A profile shows the average flexibility across the dataset. Smoothing was applied with running average windows of size 20 for (1) and 30 for (2) and (3). A tendency for higher flexibility downstream of the transcription start than upstream was found with a large transition around the start point and spikes at –25 referring to the TATA-box.

6.2.2. Twist and Roll Flexibility Profiles

Flexibility profiles of the single-step force constants were generated. Each sequence is described by the flexibility of its 494 consecutive overlapping octamers, resulting in a matrix of 624 sequences by 494 octamer positions for each flexibility measure. The average of each octamer column in a matrix is then calculated, giving a profile that is based on the alignment of the experimentally determined transcription start sites. Note that an octamer represents the nucleotide position that lies at its centre. This means that the analysed promoter length is reduced slightly (–246.5 to +246.5 instead of –250 to +250).

Initially the profiles were considered without smoothing (Figure 6.4). Two random sets, identical in size to the promoter set, were used to generate analogous random profiles, shown in blue and grey. Note that the variation in a force constant across a promoter profile is negligible in comparison to the standard deviation of the octamer population (Chapter 3 Table 3.6), no strong patterns in promoter flexibility being apparent. This was further investigated by a series of t-tests. A value of T was calculated for each position along the promoters, comparing a position's mean and standard deviation to that of the octamer population. The percentage of positions along the average promoter that are significant (magnitude of T greater than 2.58) was calculated (Table 6.1). Analogous values were obtained for one of the random profiles as a control.

Figure 6.4: *Promoter (black) and two random (blue and grey) flexibility profiles. The units of flexibility being kJmol^{-1} degrees^{-2}. a) The k^{-}_{roll} profiles. b) The k^{+}_{roll} profiles. c) The k^{-}_{twist} profiles. d) The k^{+}_{twist} profiles*

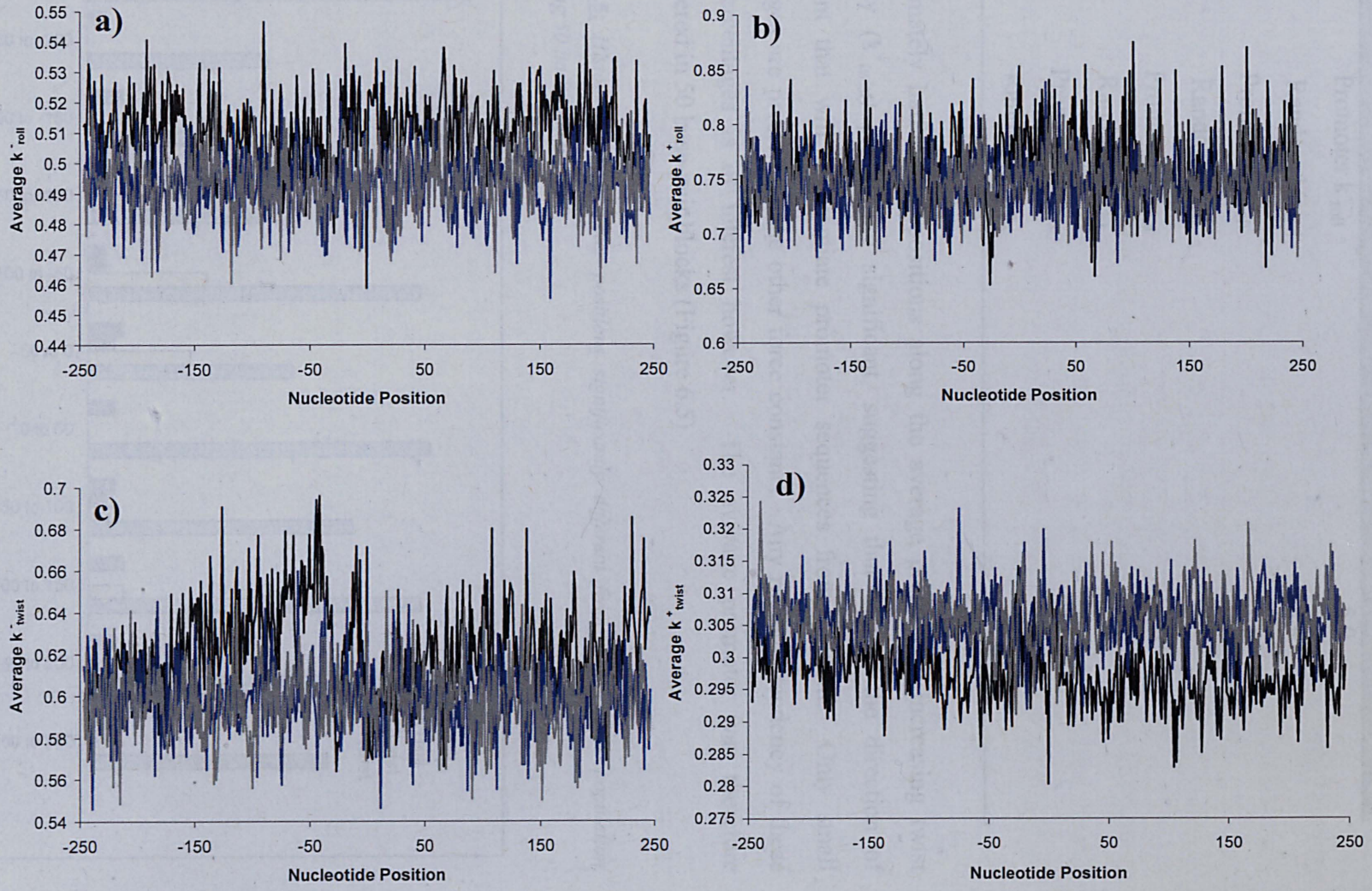
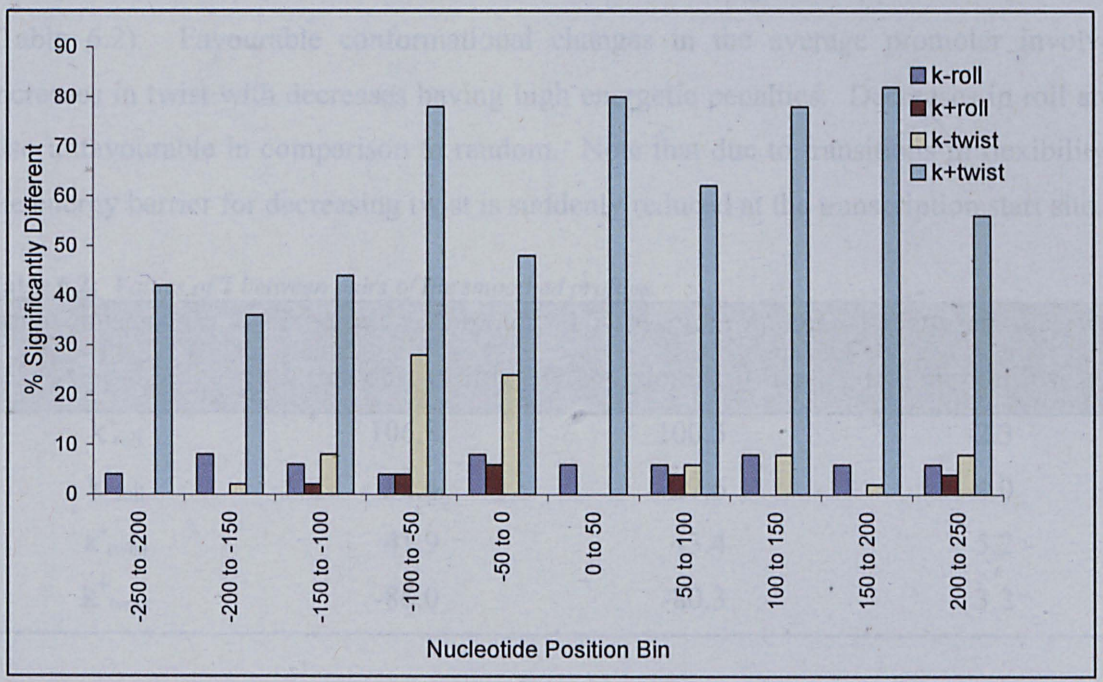


Table 6.1: *Percentage of the profiles significantly different (/T/>2.58) from the octamer population.*

Profile	Percentage of length that significant
Promoter k^-_{roll}	5.0
Random k^-_{roll}	3.4
Promoter k^+_{roll}	1.6
Random k^+_{roll}	2.1
Promoter k^-_{twist}	6.9
Random k^-_{twist}	2.1
Promoter k^+_{twist}	48.6
Random k^+_{twist}	4.0

Approximately half of the positions along the average promoter's increasing twist flexibility (k^+_{twist}) profile are significant, suggesting that this is the direction of movement that will differentiate promoter sequences from random. Only small percentages are present for the other force constants. Any regional tendency of these small percentages is of interest however. The average promoter was therefore reconsidered in 50 base-pair blocks (Figure 6.5)

Figure 6.5: *Histogram of promoter positions significantly different from the octamer population, considering 50 base-pair blocks.*



In k^-_{roll}, the significant positions are randomly dispersed across the promoter length. k^+_{roll} has the lowest overall percentage with again no apparent regional tendency. Even though the significance of k^-_{twist} is only 6.9% this level rises to approximately 25% when considering the upstream region –100 to 0, suggesting that flexibility in decreasing twist is important over this promoter location. Promoter differentiation via k^+_{twist} appears very promising.

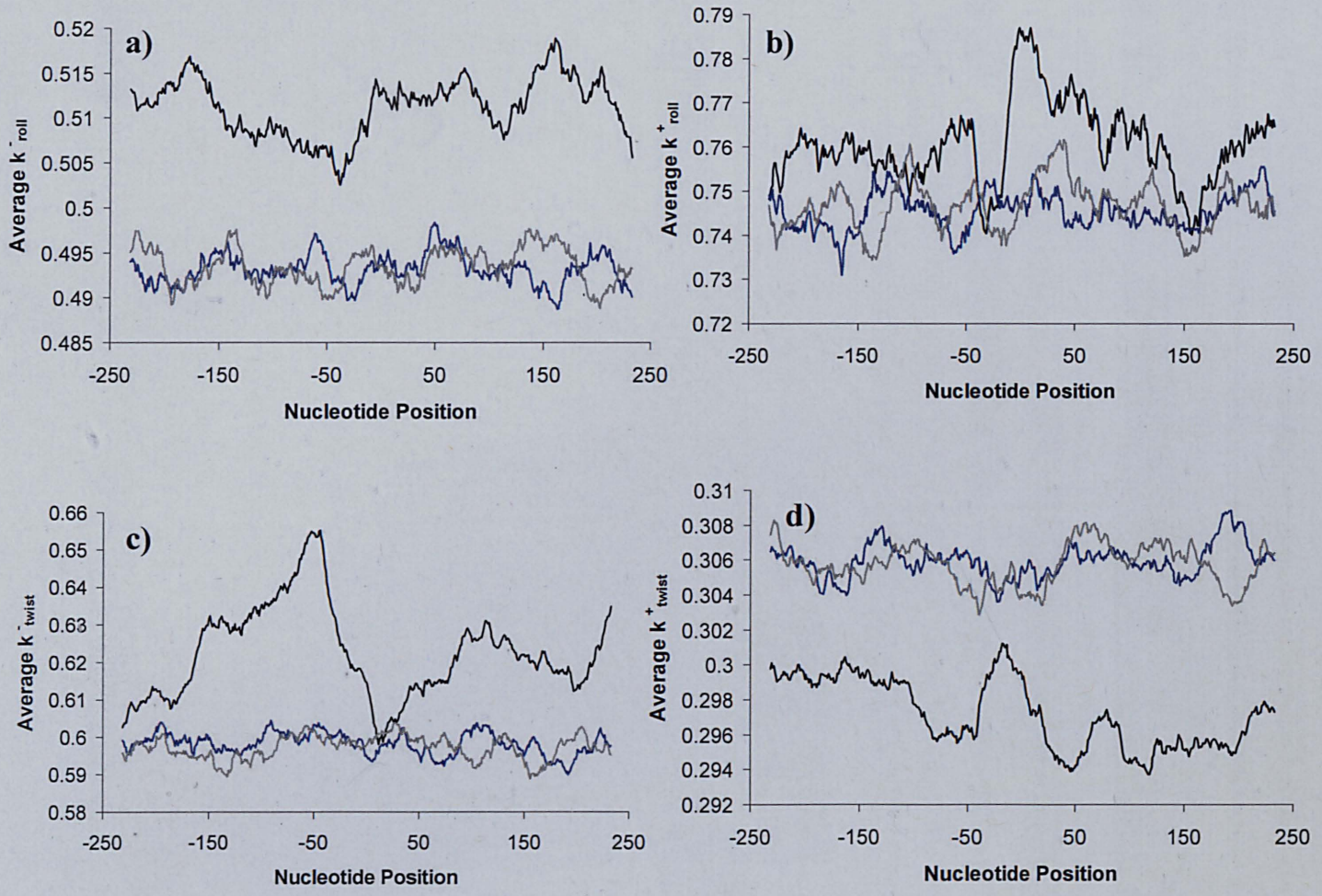
A smoothing window of 30, analogous to that used in the Location Preference and Propeller-Twist profiles (Pedersen et al., 1998), was applied to the profiles. Local transitions in flexibility were observed and compared to those of the smoothed random profiles (Figure 6.6). Remember that higher force constants refer to less flexible octamers, since more energy is required to rotate in a particular direction. Significant transitions in flexibility are only apparent in k^-_{twist} and k^+_{roll}. A clear transition from low flexibility to high flexibility is seen in k^-_{twist} between –50 and +10, supporting the idea that flexibility is greater downstream than upstream. k^+_{roll} has a transition in approximately the same place as k^-_{twist}, but in the opposite direction.

The amount of overlap between the two random profiles is far greater than the overlap of either with the promoter set. The random profiles appear collectively above the increasing twist promoter profile and collectively below the decreasing roll and twist profiles. The significance of these observations can again be quantified by T-tests (Table 6.2). Favourable conformational changes in the average promoter involve increases in twist with decreases having high energetic penalties. Decreases in roll are also unfavourable in comparison to random. Note that due to transitions in flexibility, the energy barrier for decreasing twist is suddenly reduced at the transcription start site.

Table 6.2: *Values of T between pairs of the smoothed profiles.*

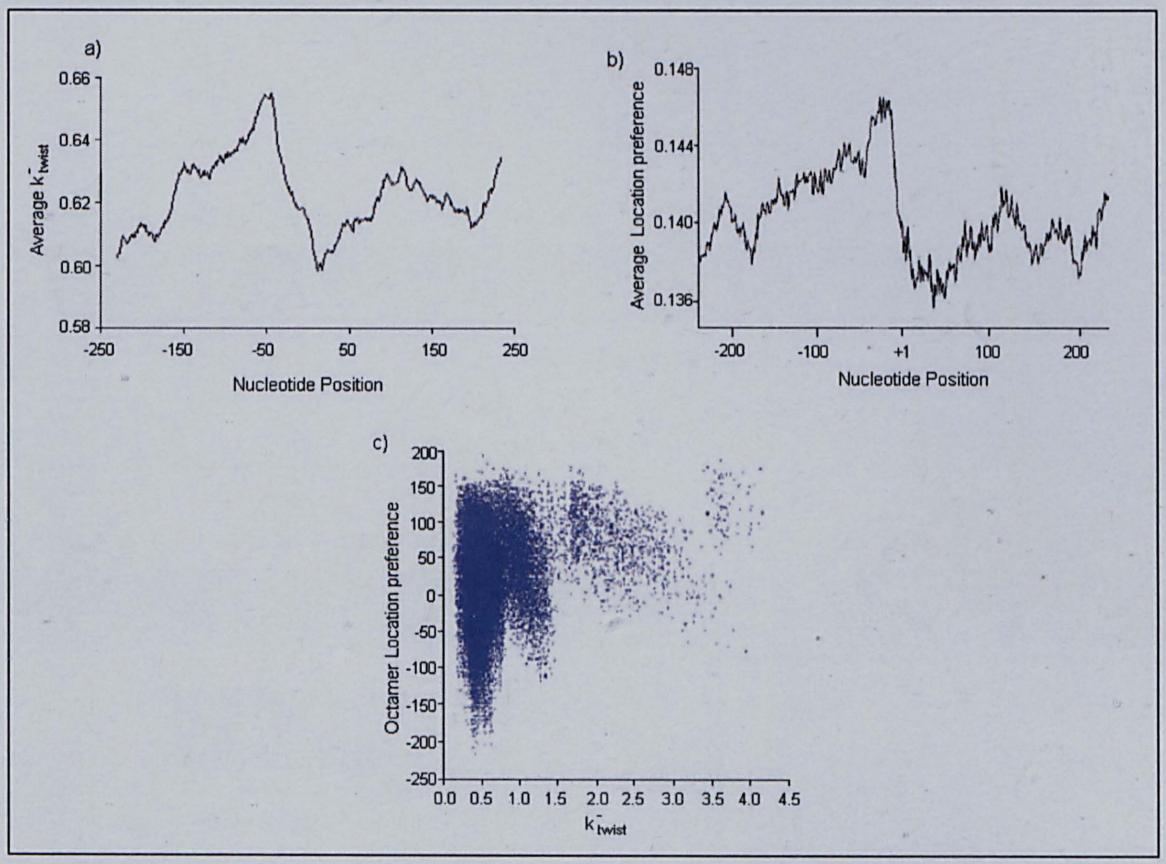
Force constant	T between promoter & random1 profile	T between promoter & random2 profile	T between the two random profiles
k^-_{roll}	106.8	100.5	-2.3
k^+_{roll}	32.4	27.6	-4.0
k^-_{twist}	41.9	43.4	5.2
k^+_{twist}	-86.0	-80.3	3.2

Figure 6.6: *Promoter (black) and two random (blue and grey) flexibility profiles with smoothing using a window size of 30. The units of flexibility being kJmol^{-1} degrees^{-2}. a) The k^{-}_{roll} profiles. b) The k^{+}_{roll} profiles. c) The k^{-}_{twist} profiles. d) The k^{+}_{twist} profiles*



A surprise finding was the remarkable similarity in the shape between the smoothed k^-_{twist} curve and the Location Preference profile (Figure 6.7a and b). This could signify a possible relationship between these two descriptors or just the octamers present in the promoter sequences. The Location Preference profile is based upon the frequencies of trinucleotides found in DNA wrapped around nucleosomes (Satchwell et al., 1986; Goodsell and Dickerson, 1994). These trimer descriptors have been converted into octamer descriptors by summing the 6 overlapping consecutive trimers that form each octamer. This allows correlations between the Location Preference and k^-_{twist} to be tested. No monotonic or linear association is apparent between k^-_{twist} and the Location Preference with a Spearman rank correlation coefficient of only 0.168 (Figure 6.7c). However, a partial dependence of k^-_{twist} on the Location Preference and vice versa can be observed. When the Location Preference is very low, k^-_{twist} is restricted to low values and when k^-_{twist} is high, the Location Preference is restricted to high values.

Figure 6.7: *Comparison of k^-_{twist} to Location Preference. a) k^-_{twist} profile. b) Location preference profile. c) Plot of k^-_{twist} versus location preference. Note that the location preference has been converted into an octamer descriptor by summing the 6 constituent trimers.*



6.2.3. Upstream versus downstream flexibility

Pedersen et al., (1998) investigated the average flexibility of an upstream region (-200 to –50) relative to a downstream region (+1 to +150) concluding that flexibility is generally higher downstream. The TATA-box region was deliberately excluded, so as not to bias results. Analogous comparisons were performed with the roll/twist flexibility force constants by summing values of consecutive overlapping octamers within these two regions and noting the number of promoters having a higher flexibility downstream. Approximately only half of the promoters are more flexible downstream than upstream, with respect to the roll and twist measurements (Table 6.3). Even when considering the coding (Table 6.4) and non-coding (Table 6.5) subsets separately, the relative regional flexibility of downstream to upstream is still around 50% and therefore is not affected by any "codon usage bias" (Pedersen et al., 1998).

Table 6.3: *Promoters having higher flexibility downstream (+1 to +150) than upstream (-200 to –50)*

Flexibility Parameter	% of promoters having higher flexibility downstream
k^-_{roll}	49
k^+_{roll}	47
k^-_{twist}	53
k^+_{twist}	55

Table 6.4: *Coding promoters having higher flexibility downstream than upstream*

Flexibility Parameter	% of coding promoters having higher flexibility downstream
k^-_{roll}	48
k^+_{roll}	45
k^-_{twist}	55
k^+_{twist}	59

Table 6.5: *Non-coding promoters having higher flexibility downstream than upstream*

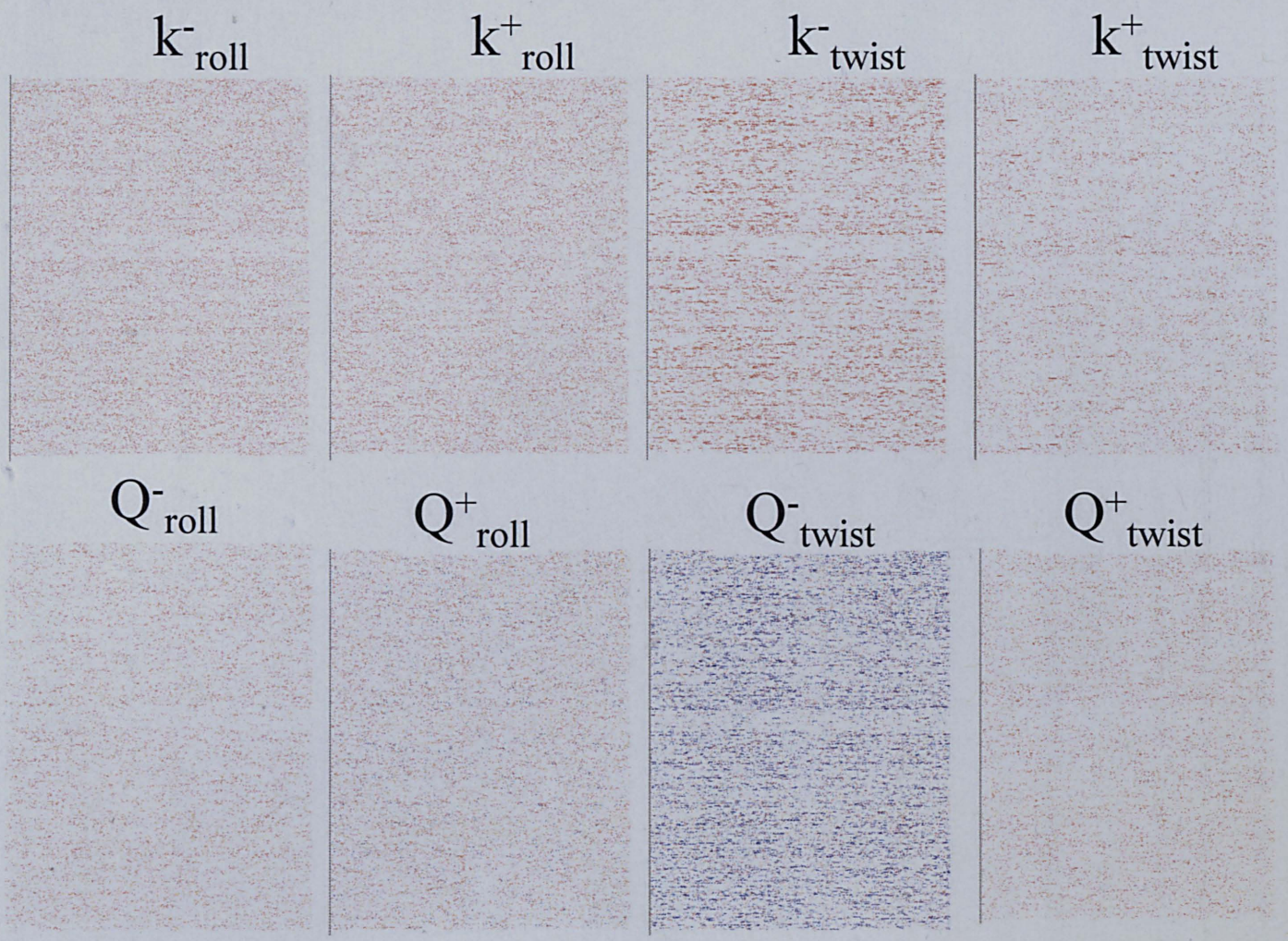
Flexibility Parameter	% of non-coding promoters having higher flexibility downstream
k^-_{roll}	50
k^+_{roll}	48
k^-_{twist}	52
k^+_{twist}	53

6.2.4. Flexibility Profiles Of Individual Promoters

The flexibility profiles that have been constructed view patterns in flexibility across the whole dataset, however they say nothing about the individual promoters and how closely they are represented by their average. Different classes of promoters may exist that cancel out one another's patterns in flexibility when their distinguishing features are combined. The comparison of single sequence profiles is therefore needed. For each promoter and flexibility parameter, a vector descriptor was calculated, the n^{th} element being the standard deviation distance of the parameter from its population mean at the n^{th} overlapping octamer position. This results in a matrix of size 494 by 624 for each flexibility parameter, where each row represents a promoter and each column an overlapping octamer position.

The 1-step flexibility parameter matrices are shown in Figure 6.8. Red means that the parameter at that point is significantly high (two standard deviations greater than the population mean) and blue that it is significantly low. The majority of the Q^{-}_{twist} matrix is blue and corresponds to the red pattern observed in the k^{-}_{twist} matrix plot. Note that blue in the partition coefficient matrices and red in the force constant matrices refer to patterns of rigidity not flexibility. Ideally the matrix plots would contain clear vertical lines, indicating flexibility trends common to all promoters. Even though this is clearly not the case, faint vertical lines through the middle are seen in the matrices associated with decreasing twist flexibility. A light horizontal band is also present, suggesting a specialised subclass of promoters. The vertical lines occur between the TATA box region and just after the start site, referring to a common highly fluctuating flexibility between –50 and +10 in the promoters. The shading appears to be slightly darker to the left of the start site than to the right, meaning that there may be a tendency for upstream to be more rigid than downstream. This would be in agreement with previous work (Pedersen et al., 1998).

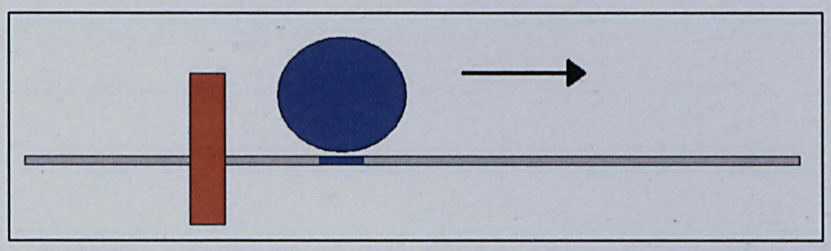
Figure 6.8: *Single step flexibility matrices*



6.3. Conclusions

Twist flexibility is the important parameter for identifying promoter sequences. Approximately half of the positions along the average promoter are significantly different from the octamer population with respect to k^{+}_{twist}. k^{-}_{twist} also appears important in the upstream region –100 to 0. After smoothing, significant transitions appear between –50 and +10 from low to high flexibility in k^{-}_{twist}. Characteristics in promoter flexibility were identified in the promoter profiles that were absent in the random profiles. It is favourable to increase the twist of a promoter, but high energetic penalties are associated with decreases in either roll or twist. Due to the transitions however, the energy barrier for decreasing twist is suddenly reduced around the transcription start site. This enables the polymerase to unwind the DNA, but in one direction only (Figure 6.9).

Figure 6.9: *The decreasing twist energy barrier (red) blocks Polymerase (blue) from travelling in one direction along the DNA (grey). The direction of transcription is restricted to that marked by the arrow.*



Relative flexibility of downstream to upstream is not affected by codon usage and neither up or downstream tends to be generally more flexible, when summing octamer values in these regions. There is a remarkable similarity between the shapes of the k^{-}_{twist} and Location Preference profiles. However no monotonic or linear association is present between them. When the Location Preference is very low k^{-}_{twist} appears to be restricted to low values and when k^{-}_{twist} is high the Location Preference appears to be restricted to high values. Note that these dependencies are directional.

Whether a protein can successfully bind to a particular piece of DNA not only depends upon how easily it is able to flex the DNA sequence, but also upon the shape of the unbound DNA. A sequence having a similar bound and unbound conformation does

not necessarily need to be flexible for protein recognition. Perhaps a more valid correlation can be found between promoter activity and a descriptor that combines the flexibility of DNA with information about its energy minima conformation (e.g. the structural probabilities introduced in Chapter 4).

Currently the application of multiple sequence profiles is restricted to pre-aligned sequences. It would be of great benefit to develop a tool that generates a multiple alignment as the first step in calculating a structural fingerprint. The remainder of this thesis therefore concentrates on developing a novel Hidden Markov Model technique that aligns DNA not by its nucleotide sequence but by its structure. In order for this new alignment tool to be successful it should account for flexibility and the fact that a single sequence can adopt one of several structures.

Chapter 7:
Hidden Markov Models

Hidden Markov Models (HMMs) have been extensively used to produce multiple alignments of, and find patterns within, sets of protein sequences (Haussler et al., 1993; Krogh et al., 1994a) and DNA sequences (Churchill, 1989; Baldi et al., 1996) sequences. Many other applications outside of bioinformatics also exist, including face image retrieval (Martinez, 1999), the study of electrocardiogram signals within the field of medicine (Koski, 1996) and its well-founded use as an analysis tool in speech recognition, dating back to the 1970's (Rabiner, 1989).

This chapter introduces the concepts behind HMMs and includes some general examples of models and their architectures. A detailed account of model construction with algorithmic solutions is presented, and successful applications from the literature to biological sequences are then given. Finally, a brief review of previous work on building structural HMMs is included, since successfully developing this style of model with the octamer database parameters is a main objective of this research.

7.1. Random variables, Markov chains & Hidden Markov Models

An HMM is a probabilistic model describing a stochastic/random process and is based upon the theory of Markov chains that was invented by Andrei A. Markov. A Markov chain is a sequence of random variables or, using the HMM terminology, a sequence of states. A detailed discussion of random variables and Markov chains with informative examples can be found in the literature (Norris, 1999; Grimmett and Stirzaker, 2001). In brief, a discrete random variable or state that has six equally probable outcomes can be used to describe a single roll of a fair dice. Continuous random variables also exist, such as that used to describe how far a ball might be thrown. In general each state in an HMM emits a value of the property being modelled with its own individual probability distribution.

It is the order of the Markov chain that defines the dependency of a state in the sequence upon its predecessors. Here we are interested in the classical 1st order Markov chain, where the next variable in a chain is solely dependent on the current variable and totally independent of all others previous to this (see Equation 7.1).

> "...conditional on its present value, the future is independent of the past"
> (Grimmett and Stirzaker, 2001)

A random walk describing the movement of an isolated particle is a clear example of this, since the position of the particle after a time interval only depends on its current position. It has no memory of the positions it has travelled through in the past.

$$P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \ldots] = P[q_t = S_j | q_{t-1} = S_i], \qquad \textbf{Equ. 7.1}$$

N.B. notation taken from Rabiner (Rabiner, 1989),

where q_t is the state at time t chosen from a set of states (S_i, S_j, S_k, ...)

Note that P[A|B] is the conditional probability of event A given that event B has already occurred. $P[A_t|B_{t-1},C_{t-2},D_{t-3}]$ is the probability of A given the sequence of events D, C then B has already happened. The above likelihood of state i being followed by state j is known as the transition probability, a_{ij} and the probability that a particular state i will emit the value k is known as an emission probability, $e_i(k)$.

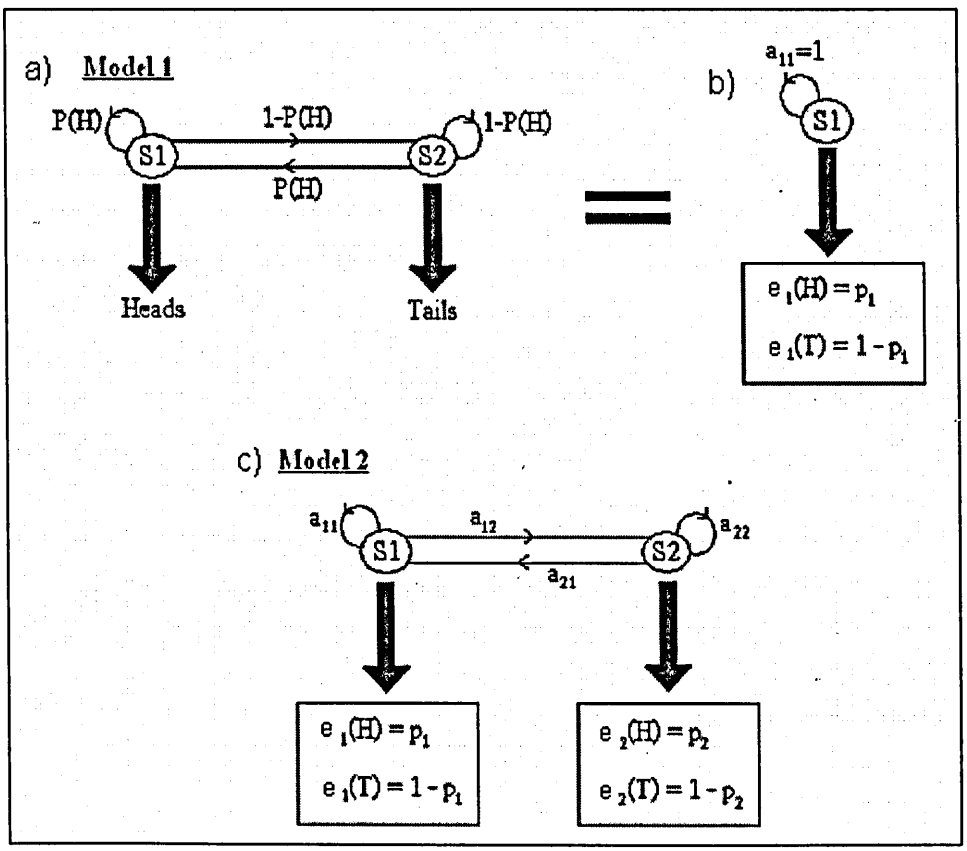
A path through a model's states, leads to a sequence of observations. Since different state paths can lead to the same observation sequence, the Markov process is hidden. This statement will be clarified by the examples of HMMs given in the next section. The information needed to completely describe an HMM can be summarised by the five parameters (Rabiner, 1989) listed below.

(1) The number of states in a model, N.
(2) The number of observables, M.
(3) The state transition probabilities, a_{ij}, defined by an N x N matrix A.
(4) The emission probabilities, $e_i(k)$, defined by an N x M matrix B.
(5) The probability of starting in each state, defined by vector C of length N.

7.2. Some simple examples of HMMs

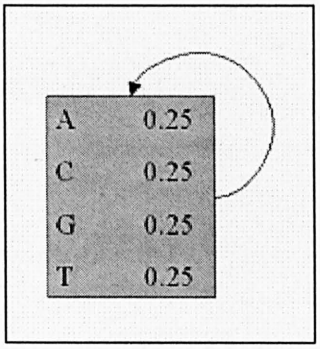
The basic ideas behind an HMM will now be clarified by three simple examples, the first being the biased coin example (Rabiner, 1989). There is a room with a barrier and on one side of the barrier some biased coins are being flipped and on the other side nothing is known but the outcome sequence of heads (H) and tails (T). Two different ways of modelling this coin system are shown in Figure 7.1. Circles denote the states, with the arrows between them representing the transitions and the block arrows pointing to the emissions. Model 1 (Figure 7.1a) consists of two states, one that only emits heads (S1) and one that only emits tails (S2) and is therefore really a single state HMM as shown in Figure 7.1b. It should be noted that a single state HMM is not really a true HMM since its state path is not hidden. Model 2 (Figure 7.1 c) also consists of two states. However this time each state represents a coin and the state path that generates the observed outcome sequence will be truly hidden.

Figure 7.1: *The biased coin example. a) Model 1 has two states (S1 and S2). S1 emits heads and S2 emits tails. The probability of being in S1 is P(H) and the probability of being in S2 is 1-P(H). b) An equivalent single state representation of model 1, showing that the state path is not hidden. c) Model 2 is a true HMM that has two states (S1 and S2), each with their own individual probabilities of emitting H or T. The likelihood of travelling from state i to j is depicted by the transition probabilities (a_{ij}).*



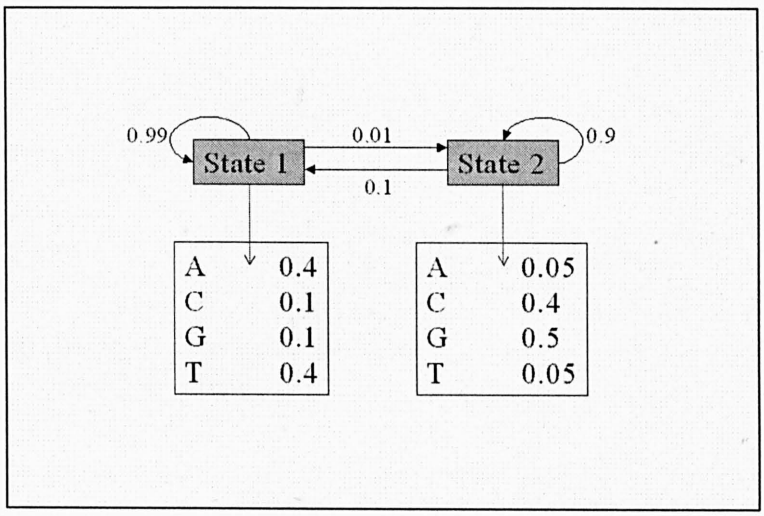
The second example is analogous to model 1 of the biased coin system, because again it is not a true HMM due to unhidden state paths. It illustrates how DNA sequences can be modelled using the discrete four-letter alphabet to allow patterns in the nucleotide sequences to be determined. The single state is shown as a rectangle with the emission probabilities contained within it (Figure 7.2). Sequences of the same length will be generated with the same probability (0.25^L, where L is the length). This important model is the null hypothesis, meaning that it is used as a standard to measure an HMM's predictive ability against. Model scoring is discussed in section 7.4.2.

Figure 7.2: *The DNA Null Hypothesis Model*



The final simple example of an HMM models DNA using two states, an AT rich state and a GC rich state (Churchill, 1989) see Figure 7.3. It can be seen from the given transition probabilities that it is far more probable to iterate back to the current state than to move into the other state, hence the model has a high probability of generating either AT rich or GC rich sequences.

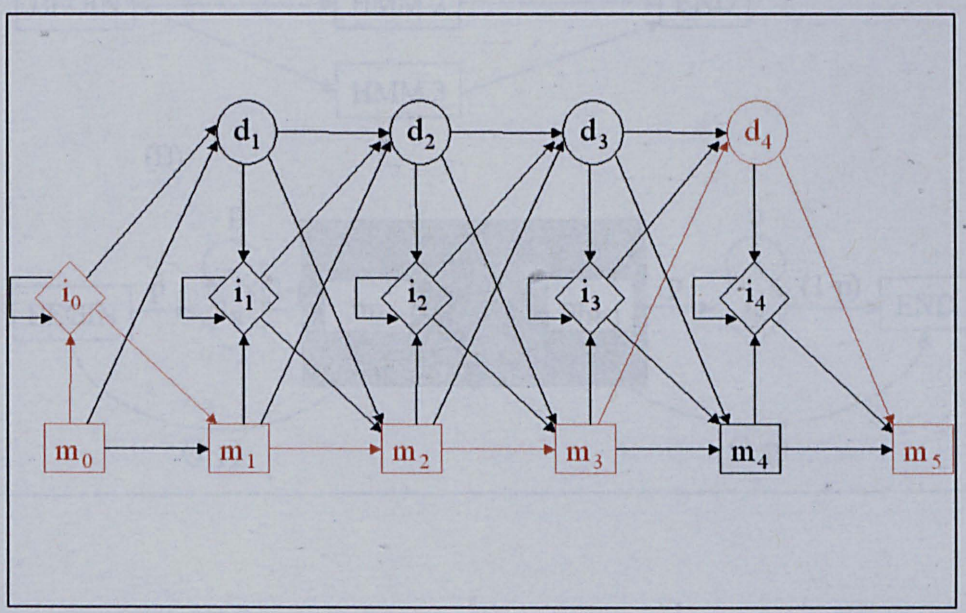
Figure 7.3: *HMM representing GC & AT rich DNA sequences (Churchill, 1989)*



7.3. Model Architecture

From the examples in the previous section it should be clear that HMMs are capable of modelling a wide variety of problems and the structure of an HMM, that is the number of states that it possesses and the way they are connected, can be tailored to suit the needs of the problem domain. This section concentrates on common architectures used to model biological sequences. Generally these architectures are based on a classical structure that is shown in Figure 7.4. Note the state types that are present. Those represented by squares are the match (m) states, which are related directly to the consensus columns in a multiple alignment. The delete (d) states, shown as circles, are used to skip a match state. For example a transition to d_2 will cause a sequence path to miss out m_2. The third type is the insert (i) state, which is used to insert extra nucleotides or amino acids between two match states, i.e. i_2 enables extra letters to be inserted between m_2 and m_3. Notice that every insert state has a recursive transition back to itself. This allows for multiple insertions between two match states. Inserts and deletes create the gaps that are commonly seen in multiple alignments. It is also usual to have a start and end state, which are null states used to depict the start and end of an observation sequence. In Figure 7.4 m_0 refers to the start state and m_5 to the end state.

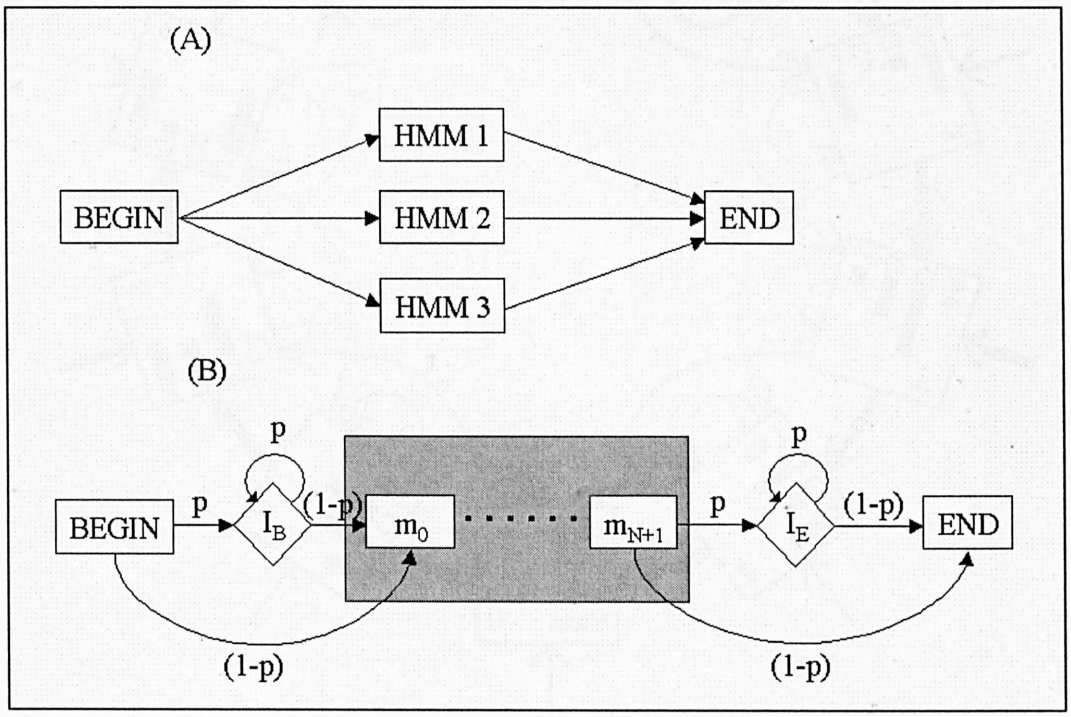
Figure 7.4: *Common HMM architecture used to model biological sequences. m_i is the i^{th} match state, i_i is the i^{th} insert and d_i is the i^{th} delete. A possible state path through the model is highlighted in red from the start state (m_0) to the end state (m_5). An observation sequence of length 4 is emitted by i_0, m_1, m_2 and m_3.*



Two architectures based upon the classical state structure are shown in Figure 7.5 (Krogh et al., 1994a). Architecture (A) connects a defined number of classical HMM structures together and was designed to identify protein sub-families. A ten-component structure of this model was applied to a set of globin sequences. The three main sub-families (alpha, beta and myoglobin) were correctly classified. A more specific example of a multi-component architecture is an HMM composed of four sub-models representing, membrane core, short turns, long loops and N or C terminus to collectively describe the transmembrane regions of β-barrel membrane proteins (Liu et al., 2003). The transitions between components of this model are however more complex than that shown in Figure 7.5.

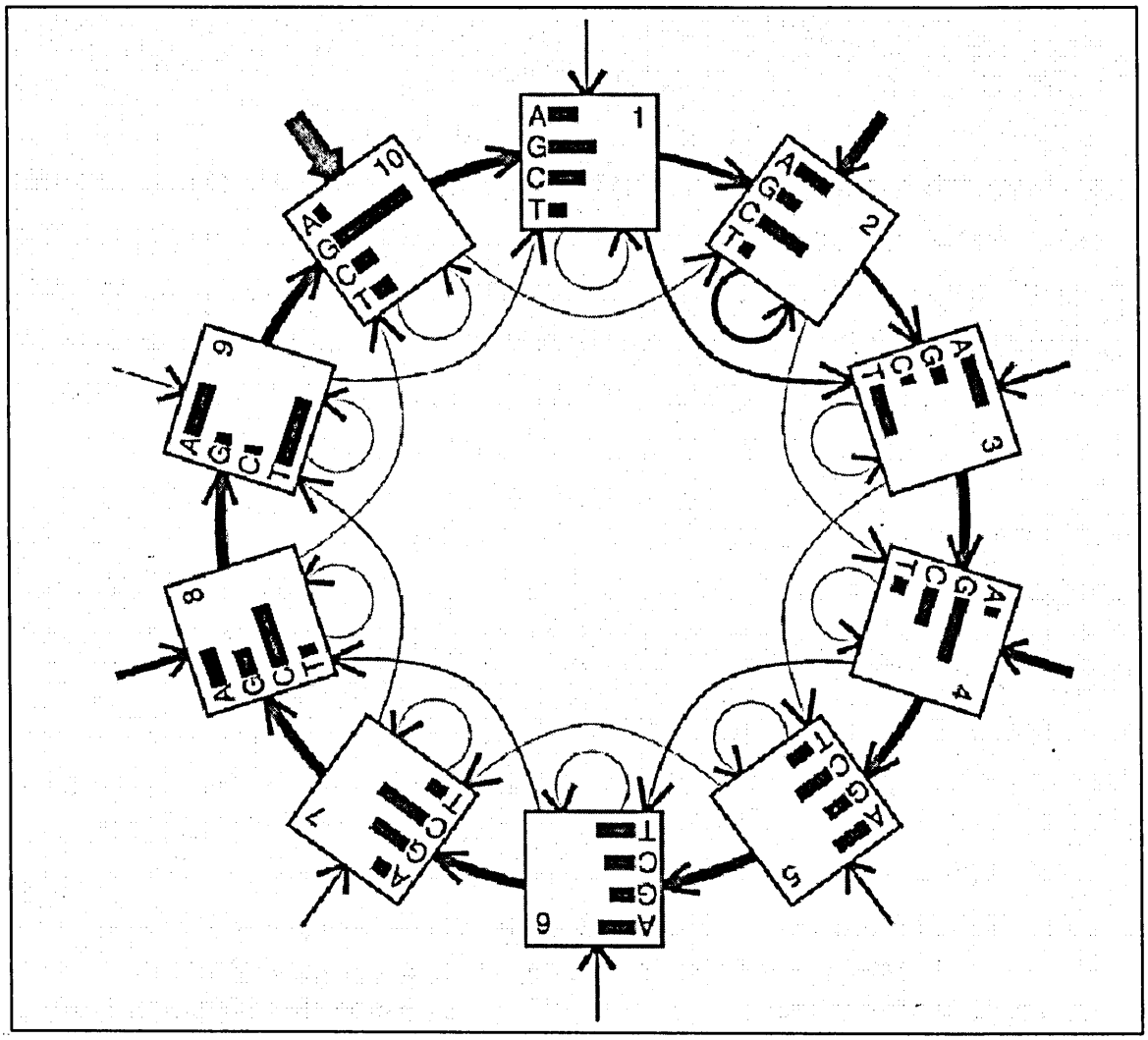
The purpose of architecture (B) is to identify domains, where a sub-sequence or several sub-sequences of a protein are important. The kinase catalytic domain was successfully modelled using 193 kinase sequences as the training set (Krogh et al., 1994a). Note that the state structure between m_0 and m_{N+1}, within the shaded box, is identical to that of Figure 7.4. For multiple domain identification, a transition from the end state back to the begin state is added to architecture (B).

Figure 7.5: *Architectures for (A)Subfamily identification, (B)Domain identification (Krogh et al., 1994a)*



An alternative wheel architecture has been used to build HMMs and multiple alignments of human exons (Baldi et al., 1995; Baldi et al., 1996). This strategy was adopted since it was thought to be suitable for determining periodic patterns in DNA (Figure 7.6). A state path can start from any of the states within the wheel. Note the absence of the traditional insert states, but instead the possibility for a match state to iterate back to itself. It was concluded from this work that strong periodic patterns that refer to nucleosome positioning signals are present within exons (Baldi et al., 1995). Confirmation is given that these patterns are not due to the secondary structure of the proteins being encoded, that they are not present to such a strong extent in introns and that they are totally absent in random sequences (Baldi et al., 1996).

Figure 7.6: *HMM Wheel Architecture (Baldi et al., 1996). The thickness of external arrows refers to the probability of starting at each state. The emissions are shown within the state boxes.*



Further examples of HMM architectures are endless and the few presented above serve as a taster to illustrate the diversity of structures that can be created. An important design factor to note is that the more complicated the architecture, the more state paths there are to explore, model parameters to optimise and the likelihood of not finding the optimal solution to the problem is increased. Model structure design is an 'art' (Durbin et al., 1998), unlike the next topic to be discussed; how to train the chosen architecture to accurately model the family of sequences being analysed.

7.4. Model construction & mathematical problems

Once the model structure has been chosen, starting values are assigned to the parameters that describe it. These values can either be randomly generated or based on intuition/prior knowledge about the specific problem domain. Next the training sequences are fitted to the starting model to produce a score. This is an overall assessment of the probability that each of the sequences was emitted from the model. An iterative procedure for adjusting the model parameters to better suit the sequences is then carried out until a reasonable converged solution has been found. The final model can then be used to generate a multiple sequence alignment of the training set and make predictions about whether further sequences belong to this set. This general method of HMM construction is illustrated by Figure 7.7. Note that within the diagram coloured boxes surrounding parts of the procedure show the stages in which three well-known algorithms, Viterbi, Forward and Baum Welch, are traditionally used.

This section presents the mathematical problems faced in model construction and analysis, introducing the algorithms that are used to overcome them. Ferguson presents these problems as falling into three categories (Rabiner, 1989).

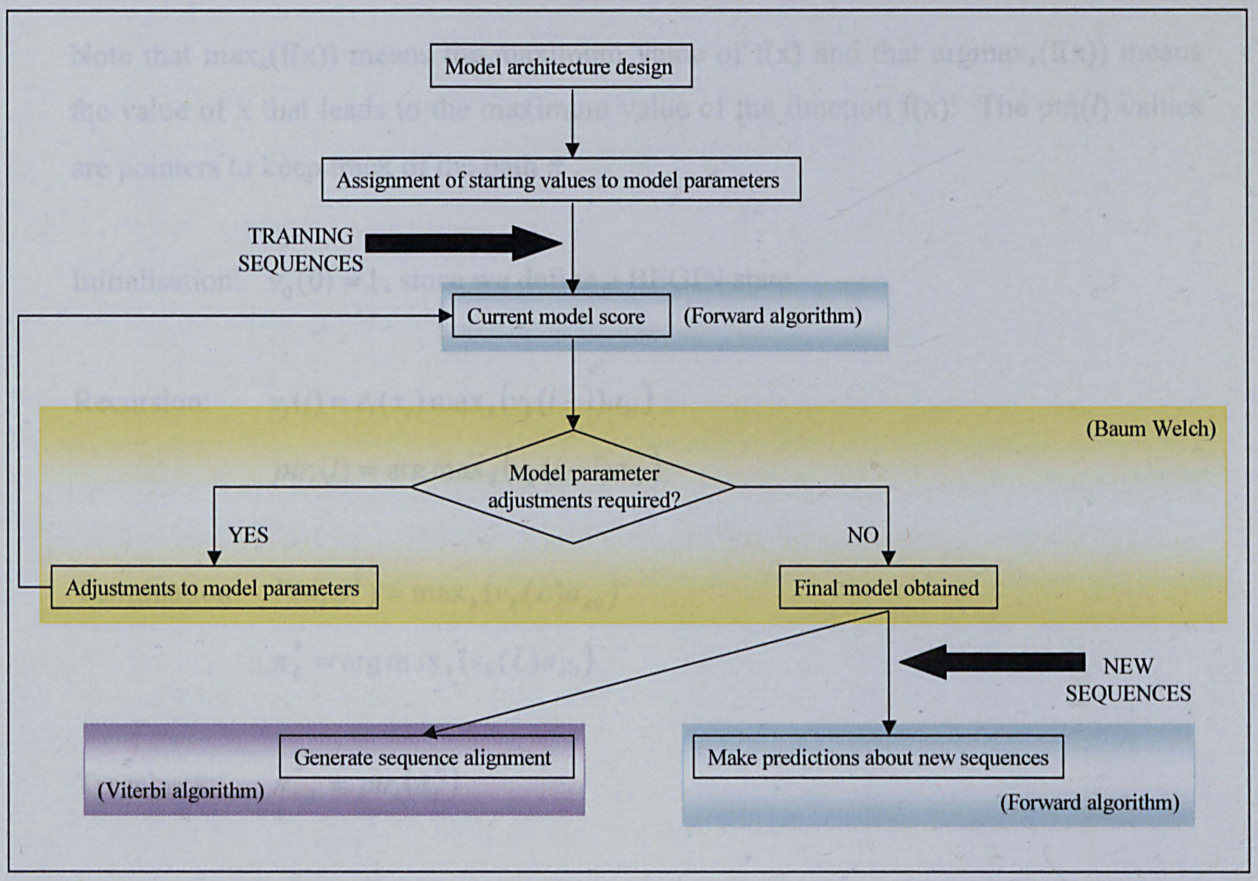
(1) Identifying the state sequence responsible for generating a particular observation sequence (approximately solved by Viterbi algorithm).

(2) Calculating the probability of generating a sequence from a given model (exact solution given by Forward algorithm).

(3) Developing a method for adjusting the model parameters so maximum performance can be obtained (Baum-Welch procedure).

These three topics shall now be discussed in turn. Excellent explanations of all the algorithmic solutions involved can be found in the literature (Rabiner, 1989; Durbin et al., 1998). Kasif and Delcher give a more general review of applying probability theory, including HMMs, to biological data (Kasif and Delcher, 1998).

Figure 7.7: *Outline of Model construction procedure. Coloured boxes show the traditional use of the three well-known algorithms, Viterbi, Forward and Baum Welch.*



7.4.1. Identifying the state path

The problem of identifying the state path of an observation sequence can only be approximately solved, due to the hidden element of HMM theory. The Viterbi algorithm is used and assumes that the responsible state path is that which has the highest probability of generating the observations. It is from these most probable paths that a multiple sequence alignment is directly obtained. The algorithm's recursive methodology follows.

x_i is the i^{th} symbol of the observation sequence.

$v_l(i)$ is probability of most probable path up to and including state l at i^{th} observation.

$e_l(x_i)$ is probability of state l emitting symbol x_i

a_{kl} is transition probability of state k being followed by state l.

L is the path length

π^* is the most probable path.

Note that $\max_x(f(x))$ means the maximum value of f(x) and that $\text{argmax}_x(f(x))$ means the value of x that leads to the maximum value of the function f(x). The $ptr_i(l)$ values are pointers to keep track of the path π^*.

Initialisation: $v_0(0) = 1$, since we define a BEGIN state

Recursion:
$$v_l(i) = e_l(x_i)\max_k\left(v_k(i-1)a_{kl}\right)$$
$$ptr_i(l) = \arg\max_k\left(v_k(i-1)a_{kl}\right)$$

Termination:
$$P(x,\pi^*) = \max_k\left(v_k(L)a_{k0}\right)$$
$$\pi_L^* = \arg\max_k\left(v_k(L)a_{k0}\right)$$

Traceback:
$$\pi_{i-1}^* = ptr_i\left(\pi_i^*\right)$$

Durbin et al give a good example of the Viterbi algorithm applied to a set of CpG islands (Durbin et al., 1998). This report uses the Churchill model (Churchill, 1989) of Figure 7.3 to illustrate how the Viterbi algorithm works. The HMM is summarised by the two matrices A and B. Remember that state 1 is AT rich, state 2 is GC rich and the probability of staying in the same state is far higher than changing to the other. It is assumed that there is an equal probability of starting in state 1 or 2.

Transition matrix A:

	State 1	State 2
State 1	0.99	0.01
State 2	0.1	0.9

Emission matrix B:

	A	C	G	T
State 1	0.4	0.1	0.1	0.4
State 2	0.05	0.4	0.5	0.05

Consider the observation sequence GCAT. The recursive Viterbi matrix (V) is shown below with any identified maxima shown in red and arrows indicating the trace back path. Elements in the first column of V, values $v_1(1)$ and $v_2(1)$, are obtained by multiplying the probability of the given state emitting G (0.1 for state 1 and 0.5 for state 2, see matrix B) by the probability of starting in that state (0.5 in both cases). Subsequently $v_1(2)$ equals the probability of state 1 emitting C multiplied by the maximum probability of reaching the state at this partcular observation time. Once the matrix is complete and the trace back has been followed it can be seen that the most probable state path is 2, 2, 1, 1 with the probability of 0.00143.

Viterbi matrix V:

$$v_l(i) = e_l(x_i)\max_k\left(v_k(i-1)a_{kl}\right)$$

	G	C	A	T
State 1	0.1X0.5 =0.05	0.1Xmax[(0.05X0.99), (0.25X0.1)] =0.00495	0.4Xmax[(0.00495X0.99), (0.09X0.1)] **=0.0036**	0.4Xmax[(0.0036X0.99), (0.00405X0.1)] **=0.0014256**
State 2	0.5X0.5 **=0.25**	0.4Xmax[(0.05X0.01), (0.25X0.9)] **=0.09**	0.05Xmax[(0.00495X0.01), (0.09X0.9)] =0.00405	0.05Xmax[(0.0036X0.01), (0.00405X0.9)] =0.00018225

7.4.2. Probability of generating a sequence from a model.

Unlike identifying the responsible state path, the probability of generating a sequence x given the model λ can be exactly solved by summing over all possible state paths π that could have generated it (Equation 7.2).

$$P(x|\lambda) = \sum_{\pi} P(x,\pi) \qquad \textbf{Equ. 7.2}$$

Note that the number of possible paths is explosive with both the number of states, N, and the sequence length, L. If it were possible for all states to be followed by all other states including themselves then the number of paths would be N^L (i.e. 10^{26} for an N of

20 and L of 20). This number is reduced to approximately $3x10^{14}$ when considering the traditional architecture of Figure 7.4. It would be infeasible to separately calculate and sum this many values. An approximate solution is to calculate the probability as that of the most probable path (Equation 7.3).

$$P(x|\lambda) = P(x,\pi^*) \qquad \textbf{Equ. 7.3}$$

Although this is often a good approximation (Durbin et al., 1998) the exact solution can be found by using the Forward algorithm. This recursive algorithm is analogous to that of Viterbi with a summation used in the recursive step instead of a maximisation, as shown below.

$f_l(i)$ is the probability of generating the sequence up to the i^{th} observation symbol with x_i being emitted from state l.

Initialisation: $f_0(0) = 1$, since we define a BEGIN state

Recursion: $f_l(i) = e_l(x_i)\sum_k f_k(i-1)a_{kl}$

Termination: $P(x|\lambda) = \sum_k f_k(L)a_{k0}$

Applying the forward algorithm to the Churchill model results in the following matrix and determination of $P(x|\lambda)$ when sequence x equals GCAT.

Forward matrix F:

$f_l(i) = e_l(x_i)\sum_k f_k(i-1)a_{kl}$

	G	C	A	T
State 1	0.1X0.5 =0.05	0.1X[(0.05X0.99)+ (0.25X0.1)] =0.00745	0.4X[(0.00745X0.99) +. (0.0902X0.1)] =0.0065582	0.4X[(0.00656X0.99)+ (0.00406X0.1)] =0.0027595562
State 2	0.5X0.5 =0.25	0.4X[(0.05X0.01)+ (0.25X0.9)] =0.0902	0.05X[(0.00745X0.01)+ (0.0902X0.9)] =0.004062725	0.05X[(0.00656X0.01)+ (0.00406X0.9)] =0.000186101725

Summing the final column in the above matrix gives the probability that the Churchill model produces GCAT, $P(x=GCAT|\lambda)$, of 0.00295. If the Viterbi approximation (Equation 7.3) was assumed and only the most probable path was considered then $P(x=GCAT|\lambda)$ would be reduced to 0.00143, an underestimation of more than 50%.

The probability that a model will generate a sequence gives a measure of how related that sequence is to the model or to the sequence family upon which the model was built. This measure must however be standardised before a fair estimation or prediction of a sequence's behaviour can be made. The standardisation involves a null hypothesis, previously illustrated for DNA sequences in Figure 7.2. The null hypothesis can be thought of as a random sequence generator that for DNA typically generates a sequence with the probability of 0.25^L, L being the sequence length. Variations on protein null models have been discussed (Barrett et al., 1997) and include a flat distribution of amino acid frequencies, a distribution to represent frequencies within the entire population of all proteins and occurrence counts of amino acids within training set sequences. The score given to a sequence x, S(x), is commonly calculated as the log odds ratio shown in Equation 7.4.

$$S(x) = \log_2\left(\frac{P(x|\lambda)}{P(x|null)}\right) \qquad \textbf{Equ. 7.4}$$

An S(x) of zero means that sequence x is equally likely to belong to the model λ as it is to a totally random model. A positive score indicates that the sequence is more likely than random chance to belong to the model and a negative score less so. Note that the logarithm is taken to the base 2, which gives a score measured in bits. It is very important to normalise the scores by sequence length, which the log odds ratio automatically does.

The log odds ratio score for GCAT and the Churchill model is -0.405. This is understandable, since the Churchill model has been built to favour GC or AT rich sequences and to disfavour switches between. GCGC yields a positive score of 1.96. Another way to interpret these results is to say that GCGC is 5.15 times more likely to belong to the Churchill model than GCAT, that is $2^{S(GCGC) - S(GCAT)}$.

A similar scoring method to the log odds ratio is used in the construction and optimisation of an HMM, where adjustments are iteratively made to the model parameters until the model score converges (section 7.4.3). A model should be scored by how well it fits the data it is built upon. This can be measured by calculating the average score of the training set sequences or by the log likelihood of the model, where x^j is the j^{th} training set sequence and θ is the current model's parameters.

$$\sum_{j=1}^{n} \log_2 P(x^j|\theta)$$

A suggestion to take the number of free parameters into account when comparing models has been made; score – ½ k log(n), where k is the number of parameters and n the sequence length (Churchill, 1992).

7.4.3. Adjustments to model parameters

The standard iterative gradient descent procedure for determining the set of model parameters, θ, that maximises a model's performance is Baum Welch. Adjustments are made to the emission and transition probabilities to increase the log likelihood of the model. The iterations are continued until the change in model performance is less than a chosen threshold or a maximum number of cycles have been reached. The starting parameters of a model can be chosen randomly, set to frequencies representing the entire population of biological sequences or based on some prior knowledge. Prior knowledge can be built in via PAM matrices or Dirichlet mixtures (see sections 7.4.4 and 8.3). The parameters are then adjusted by calculating the expected number of times the training set sequences use the transition from state k to state l, A_{kl}, and the emission of b from state k, $E_k(b)$. These two expectation counts are based upon the probability that k is the i^{th} state of the hidden state sequence, $P(\pi_i=k|x,\theta)$, calculated using the forward algorithm and its reverse analogue, the backward algorithm. This process shall be explained shortly, after the following questions have been answered. How are A_{kl} and $E_k(b)$ calculated from $P(\pi_i=k|x,\theta)$ and how are they used to adjust the model parameters?

The value of $E_k(b)$ across the training set is obtained by summing the mentioned probability value across all of the sequences and all of their lengths (Equation 7.5). The sequence number is j with x^j being the j^{th} sequence. The position along the sequence is i with x_i being the letter emitted at that time. A_{kl} is calculated in a similar manner to $E_k(b)$. Note though that it has to account for the two state path elements that make up the transition of interest (Equation 7.6).

$$E_k(b) = \sum_j \sum_i P(\pi_i = k | x^j, \theta) \qquad \textbf{Equ. 7.5}$$

$$A_{kl} = \sum_j \sum_i P(\pi_i = k, \pi_{i+1} = l | x^j, \theta) \qquad \textbf{Equ. 7.6}$$

Once all values of A_{kl} and $E_k(b)$ have been calculated, the new parameter set (a_{kl} and $e_k(b)$) is obtained (Equations 7.7a and b). The difference in the log likelihood of the new model and previous model is then examined, followed by repetition of the optimisation procedure using the new parameter set.

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \qquad \textbf{Equ. 7.7a}$$

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')} \qquad \textbf{Equ. 7.7b}$$

Returning to the probability calculations, $P(\pi_i=k|x,\theta)$ is related to the forward and backward variables, $f_k(i)$ and $b_k(i)$ respectively, as shown in Equation 7.8a. Equation 7.8b is used when the probability of a transition between two states at time i is required. Note that in this second equation the index of the backwards variable differs from that of the forwards variable.

$$P(\pi_i = k | x, \theta) = \frac{f_k(i) b_k(i)}{P(x|\lambda, \theta)} \qquad \textbf{Equ. 7.8a}$$

$$P(\pi_i = k, \pi_{i+1} = l | x, \theta) = \frac{f_k(i) b_l(i+1)}{P(x|\lambda, \theta)} a_{kl} e_l(x^j_{i+1}) \qquad \textbf{Equ. 7.8b}$$

The forward algorithm was presented in section 7.4.2. The backward algorithm, as opposed to the forward algorithm, starts at the end of the sequence and works backwards, recursively calculating the probability of being in state k given all the possible paths after it and till the end of the sequence. This procedure and the value of $b_k(i)$ are given below.

Initialisation: $b_k(L) = a_{k0}$

Recursion: $b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1)$

Termination: $P(x|\lambda) = \sum a_{0l} e_l(x_1) b_l(1)$

Finally substitution of Equation 7.8a into 7.5 and 7.8b into 7.6 results in the following formulae for obtaining $E_k(b)$ and A_{kl} in terms of the forward and backward variables and the current parameter set.

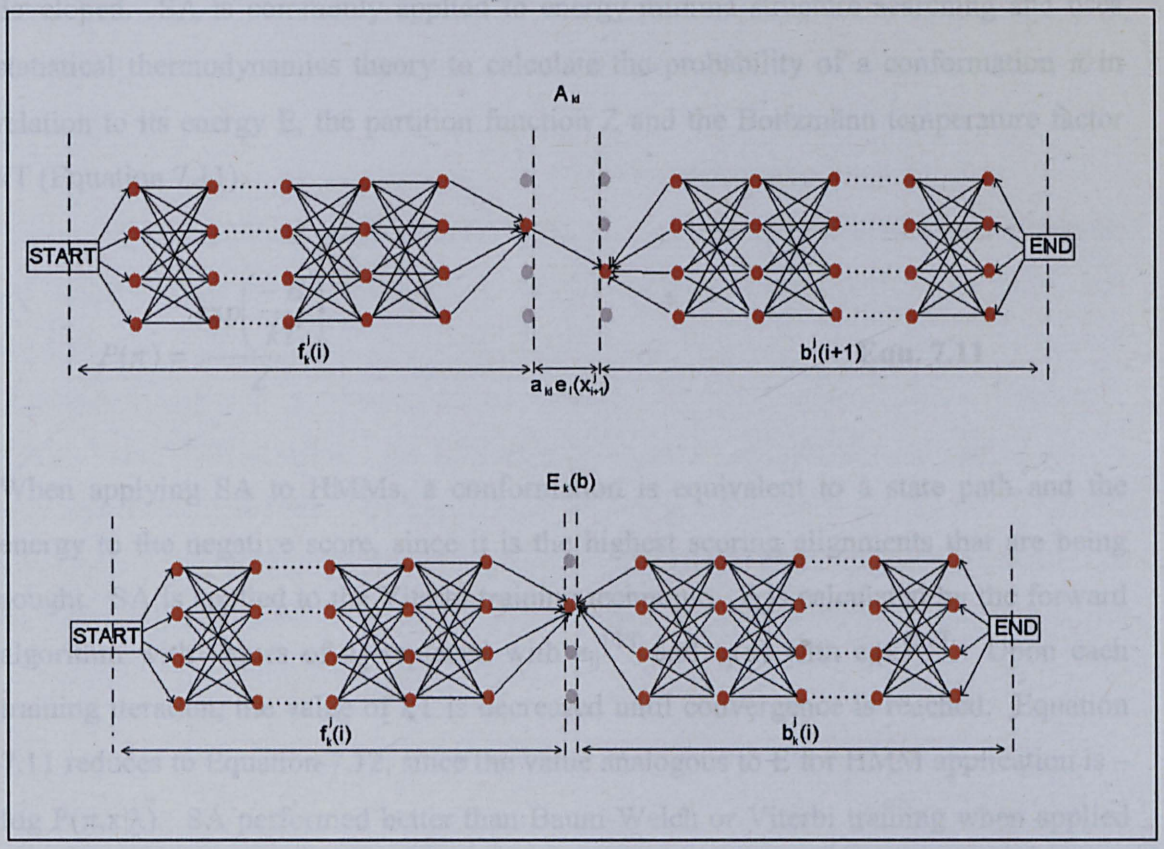
$$E_k(b) = \sum_j \left[\frac{1}{P(x^j|\lambda,\theta)} \sum_i f_k^j(i) b_k^j(i) \right] \qquad \textbf{Equ. 7.9}$$

$$A_{kl} = \sum_j \left[\frac{1}{P(x^j|\lambda,\theta)} \sum_i (f_k^j(i) b_l^j(i+1) a_{kl} e_l(x_{i+1}^j)) \right] \qquad \textbf{Equ. 7.10}$$

For a visual representation of how the forward and backward algorithms and their variables are related to the values of A_{kl} and $E_k(b)$ see Figure 7.8. The circles represent the states, those in red being involved in the calculation. The components of A_{kl} and $E_k(b)$ with their associated segments of the model are labelled across the bottom of each diagram in the figure. The reason for the index difference of the backward variable between the two expectation counts should be clearly seen.

A common alternative to the Baum Welch procedure is Viterbi training. This method has been successfully used and is often favoured due to its computational speed (Haussler et al., 1993). It is however based upon the Viterbi approximation, so the accuracy of results are compromised.

Figure 7.8: *Using the forward and backward variables to obtain the expectation counts, A_{kl} and $E_k(b)$ for a hypothetical 4 state model where transitions between all states and themselves are allowed. States are represented by circles and transitions as lines with sequence position i increasing from left to right.*



7.4.4. Further suggestions for model performance optimisation

HMMs are commonly used to obtain a multiple sequence alignment from scratch with several local optima present within the parameter space that give rise to different alignment solutions. The algorithms previously introduced have a tendency to get stuck in local optima and therefore may not find the best answer to the problem. If an alignment of a dataset is already available then using it as a starting point to build an HMM greatly reduces the parameter search space. Parameter estimation can be carried out by converting the observed counts of both symbol emissions and state transitions from the provided alignment into probabilities (Krogh, 1998). Even just a small pre-aligned subset of the training data can be useful (Krogh et al., 1994a). It is also worth considering more constrained model architectures with less deletes, inserts and therefore model parameters than the traditional model.

Extensions to the traditional training algorithms have been suggested in order to try and solve local minima problems. A simulated annealing (SA) technique (Eddy, 1995) based upon the "noise-injection" method (Haussler et al., 1993) has been developed. SA is commonly applied to energy minima structure searching and uses statistical thermodynamics theory to calculate the probability of a conformation π in relation to its energy E, the partition function Z and the Boltzmann temperature factor kT (Equation 7.11).

$$P(\pi) = \frac{\exp\left(\frac{-E}{kT}\right)}{Z} \qquad \textbf{Equ. 7.11}$$

When applying SA to HMMs, a conformation is equivalent to a state path and the energy to the negative score, since it is the highest scoring alignments that are being sought. SA is applied to the Viterbi training technique. Z is calculated by the forward algorithm with values of a_{ij} replaced with $a_{ij}^{1/kT}$ and $e_j(x)$ with $e_j(x)^{1/kT}$. Upon each training iteration, the value of kT is decreased until convergence is reached. Equation 7.11 reduces to Equation 7.12, since the value analogous to E for HMM application is $-\log P(\pi,x|\lambda)$. SA performed better than Baum-Welch or Viterbi training when applied to 10 different protein datasets (Eddy, 1995).

$$P(\pi) = \frac{P(\pi, x|\lambda)^{1/kT}}{\sum_{\pi'} P(\pi', x|\lambda)^{1/kT}} \qquad \textbf{Equ. 7.12}$$

An extension to Baum Welch and SA that involves Particle Swarm optimisation (PSO) combined with an evolutionary algorithm (EA) has been found to improve certain alignments (Rasmussen and Krink, 2003). This algorithmic hybrid (PSO-EM) is referred to here as an extension, since Baum Welch and SA solutions are used as starting points. PSO is based upon the movement of animal swarms in nature with a number of particles moving around the search space by iteratively updating their positions and velocities. The current position vector of each particle represents a possible solution to the problem. In PSO-EM the particles can also breed, in order to evolve a new generation of solutions.

Model surgery can be used to determine a better model length (the number of match states) by observing patterns in transitions to insert and delete states within each model building iteration (Haussler et al., 1993). If the fraction of optimal sequence paths that choose d_k exceeds a threshold value, γ_{del}, then that position is removed from the model. If more than γ_{ins} choose i_k then a number of new match states are inserted at position k.

An alternative to training an HMM to maximise the log odds ratio and to instead maximise the quality of an alignment, measured by the sum of pairs score (SoP), has been proposed (Rasmussen and Krink, 2003). See Equation 7.13, where l_i is aligned sequence i and D is a distance matrix, such as BLOSUM (Henikoff and Henikoff, 1992). This score considers the similarity between all pairs of aligned sequences. A gap cost is calculated for every gap in each sequence and subtracted from the SoP score (Equation 7.14). GOP is the fixed gap opening penalty, n is the gap length and GEP is the gap extension penalty. A GOP of 11 and GEP of 2 have been previously used for protein work (Rasmussen and Krink, 2003).

$$SoP = \sum_{i=1}^{n-1} \sum_{j-i+1}^{n} D(l_i, l_j), \qquad \textbf{Equ. 7.13}$$

$$GapCost = GOP + nGEP \qquad \textbf{Equ. 7.14}$$

The HMM training data should be chosen carefully. Both quantity and diversity are important factors. For protein family recognition more than 100 sequences should be used (Eddy, 1996) otherwise it is likely that the model will not have enough information to identify a pattern. When the amount of training data is limited to under this amount it is strongly advisable to add prior information into the model, commonly performed via pseudocounts, PAM matrices, BLOSUM matrices or Dirichlet Mixtures, which will all be briefly introduced shortly. One of these techniques should also be used when a larger sample of data does not have a fair spread in diversity. In such skewed sets it is sensible to apply weights to the sequences, giving less importance to those that are over represented.

A Maximum Discrimination HMM method has been developed that optimises the correct classification of sequences (Eddy et al., 1995). This procedure is effectively equivalent to using a sequence weight scheme proportional to the probability of misclassifying a sequence. A model constructed via this method is capable of recognising more than one sequence family from information given in the training data. However this can be a disadvantage, since the effect of a single false positive in the training set will be greatly increased. Several variations of the maximum discrimination weighting scheme have been explored (Karchin and Hughey, 1998) and other methods exist that use tree based clustering of sequence similarity (Gerstein et al., 1994).

Pseudocounts prevent an amino acid or nucleotide from being given a probability of zero in an alignment column. This is essential because if a sequence fits well to the majority of a model, but has a single zero probability then, due to calculating products, the entire sequence will be given a probability of zero. Pseudocounts work by adding a constant to each observed amino acid count and then renormalizing the distribution. An analogy has been made between this method and a single component Dirichlet mixture (Sjolander et al., 1996).

PAM (Dayhoff et al., 1978) and BLOSUM (Henikoff and Henikoff, 1992) matrices are amino acid substitution matrices. They tabulate the probability of replacing one amino acid with another and therefore define the similarity between pairs of amino acids. Substitution matrices form a vast research area in themselves. For further discussion see Chapter 8.

A Dirichlet Mixture is made up of several components or densities. The mixture assigns differing weights to each component, which are called the mixture's coefficients. A component is a probability density over a set of probability vectors, each vector describing an amino acid distribution. The use of several components means that different relationships between the amino acids can be considered within one alignment. For example, one component might favour that certain residues be buried whilst another may favour solvent exposed residues. More detail about Dirichlet mixtures can be found elsewhere (Sjolander et al., 1996; Durbin et al., 1998).

7.5. Applications of HMMs to biological sequences

Three applications of HMMs to biological sequences were presented in section 7.3: the recognition of protein family and sub-family sequences (Krogh et al., 1994a), domain identification (Krogh et al., 1994a) and examining periodic patterns within the exons of Human DNA (Baldi et al., 1995; Baldi et al., 1996). This section includes some further examples of applications to biological sequences.

A classic example of successful protein family recognition is an HMM built from 400 randomly selected globin sequences that was used to make predictions about a further 225 globins and 19,458 non-globins (Haussler et al., 1993). Viterbi training rather than Baum Welch was used and model surgery performed to optimise the model length. The results obtained were very promising with very low numbers of false positives and false negatives. Excellent agreement of the final HMM's sequence alignment with a previously obtained structural alignment was also observed. Further examples of family classification can be found (Baldi et al., 1994; Karplus et al., 1998). The latter example is a different approach to HMMs, designed to detect remote homologies of a single target sequence. It has been shown to be three times as effective as pairwise methods (Park et al., 1998). Although successful, the globin recognition problem is not necessarily best thought of as a pure classification technique. A characteristic motif may have several overlapping occurrences within a single sequence. Therefore instead of finding only the single best alignment of a sequence to a model it is wiser to consider several high scoring alignments (Bucher et al., 1996).

As well as performing sub-family recognition on a protein set the opposite procedure of putting sub-family members back into their parent families can be carried out. HMMs classifying 47 of 60 glycosyltransferase families back into just four superfamilies have been constructed (Kikuchi et al., 2003). From the results obtained useful predictions about the evolutionary history of the original families were made. A model has been constructed to predict whether peptides bind to certain cell surface marker molecules associated with the immune system (Kato et al., 2003), illustrating that HMMs are clearly suited to any sort of classification problem not just family recognition.

HMMs can be used to locate genes. E. Coli DNA has been analysed using a composite HMM structure that is comprised of codon triplet models, states representing intergenic regions, codon start positions and stop positions (Krogh et al., 1994b). 80% of the genes were accurately predicted and 90% had their approximate position correctly found. False positives were often found to refer to known protein sequences, suggesting that they may be correctly identified but undiscovered genes. Two further refinements have been made to the above procedure (Krogh, 1997), the use of Class HMMs (CHMMs) and determination of the most probable gene prediction instead of state sequence. CHMMs are trained to optimise recognition rather than model statistics from the training sequences (Krogh, 1994c) and involve assigning class labels to the sequences. These labels ('C' for coding, 'I' for intron and 'O' for intergenic) are emitted from each state along with a nucleotide. Every allowed path through a CHMM must have a state label sequence that agrees with the observed labels. Further attempts to apply Markov Models to the problem of locating genes include the use of a Markov Chains/Bayes method, GeneMark (Borodovsky and McIninch, 1993) and then later an HMM version of GeneMark, GeneMark.hmm (Lukashin and Borodovsky, 1998).

Another strategy for finding genes is to search for promoter sequences. The consensus sequence of RpoD promoters in Campylobacter jejuni was investigated with an HMM trained to identify motifs upstream of known genes (Petersen et al., 2003). Research has also been carried out on locating other protein binding sites across genomes, for example the Integration Host Factor (IHF) and Factor for inversion stimulation (FIS) within the E.Coli genome (Ussery et al., 2001). A binding model to denote the IHF/FIS sites was placed between two background states that both represent the nucleotide composition within the entire genome. The transition probability from the 1st background state to the binding model is related to the posterior probability of finding a site within the genome. Several occurrences of the sites can be searched for within E.Coli by adding a transition between the two background states. This is analogous to the domain identification architecture discussed in section 7.3 (Krogh et al., 1994a).

HMMs have been built to recognise splicing sites (Yin and Wang, 2001). Splicing is the removal of introns and rejoining of exons in mRNA. The splicing sites can be either donor or acceptor sites and are the positions for intron removal and RNA rejoining. Four separate HMMs were built, true donor, false donor, true acceptor and false acceptor models. A sequence is then used as input to all the models and an acceptor score and donor score is obtained as ratios of the related true and false models. The donor model had 9 states with state 4 only emitting G and state 5 only emitting T. The acceptor model had 16 states with state 14 only emitting A and state 15 only emitting G. 92% of the true donor sites and 91.5% of the true acceptor sites were correctly identified (Yin and Wang, 2001).

Publicly accessible databases of HMMs have been generated, Pfam (Bateman et al., 2000), SUPERFAMILY (Gough et al., 2001) and PANTHER (Thomas et al., 2003). A major advantage of having an HMM library is that a new sequence can be automatically classified. The growth of sequences within these databases that attempt to represent all existing sequences is illustrated by the change in the number of families represented by Pfam. In 2000 Pfam contained 1815 families (Bateman et al., 2000) and 4 years later it contained over 6190 (Bateman et al., 2004).

7.6. Building Structural information into HMMs

Sequence and structure information can be combined to enhance the predictive ability of models. A known structural alignment can be used as a starting solution for sequence alignment (Al-Lazikani et al., 2001) or sequence can be used to identify structural characteristics. The location and orientation of alpha helices in transmembrane proteins has been predicted using ten-fold cross validation and seven state types to represent different residue types, e.g. helix loops (Sonnhammer et al., 1998). These models were further developed using sequence labels (M for membrane, I for inside/cytoplasmic) to form CHMMs (Krogh et al., 2001). Transmembrane models have also been constructed based on differences between amino acid distributions in various structural parts (Tusnady and Simon, 1998). The five states, inside loop (I), inside tail (i), membrane helix (h), outside tail (o) and outside loop (O) were used with specialised transitions to reflect known structural characteristics, such as a tail coming

after a helix can be followed by another tail or a loop, but only on the same side of the membrane. HMMs encoding structural characteristics with conditional probabilities that an amino acid belongs to one of thirteen structural types (i.e. the probability of Alanine being in a loop or coil) have also been used (Stultz et al., 1993; White et al., 1994).

Improved discrimination between CAP binding and non-binding sites of DNA was found after the addition of structural information to an HMM (Thayer and Beveridge, 2002). Roll/tilt bending dials that measure the probabilities of dinucleotides having a range of geometries were used to describe the dynamical structure of DNA. From these dials two types of probabilistic outcomes were calculated, firstly the probability that the base pair step α has the geometry k,l ($P_\alpha(k,l)$) and secondly the probability that a particular geometry is due to a certain base pair step ($P_{k,l}(\alpha)$). Initially an HMM is built upon sequence alone, with 10 observation symbols describing the 10 distinct dinucleotides. The structural HMM is then formed by combining the two previously mentioned probabilistic outcomes with the emission probabilities of the sequence based HMM. Note that the transition probabilities of the original HMM remain unchanged and that both model parameter optimisation and alignment is based purely upon patterns in the dinucleotide sequence.

Two stages are involved in merging structure with sequence. First the probability of state i emitting dinucleotide α ($e_{i\alpha}$) is translated into the probability of state i emitting geometry k,l (e'_{ikl}), see Equation 7.15.

$$e'_{ikl} = \sum_{\alpha} [e_{i\alpha} P_\alpha(k,l)] \qquad \textbf{Equ. 7.15}$$

Then e'_{ikl} is translated back to the probability of state i emitting dinuceotide α but this time with the inclusion of structural information ($e''_{i\alpha}$), see Equation 7.16.

$$e''_{i\alpha} = \sum_{k,l} [e'_{ikl} P_{k,l}(\alpha)] \qquad \textbf{Equ. 7.16}$$

As well as a noticed improvement to the recognition of binding sites, a further improvement was also observed when the HMM was restricted to sequence only in a highly conserved 5 base pair consensus region (Thayer and Beveridge, 2002).

HMMs based upon the secondary structure of proteins have been built using a 3-letter alphabet for the alignment of the key structural characteristics, helix (H), strand (E) and coil (C) (DiFrancesco et al., 1997; Di Francesco et al., 1999). A series of models, each representing a different topology, were constructed and stored in a database called FORESST (Di Francesco et al., 1999). Successful rates of test set predictions confirm the validity of this novel procedure (DiFrancesco et al., 1997; Di Francesco et al., 1999; Holbrook et al., 1999). The reduction in size from the 20-letter amino acid alphabet to this 3-letter structural alphabet means that the number of learning parameters is greatly reduced with the requirement of less training data.

It is difficult to identify the most important structural descriptor for a set of sequences, since the most discriminating factor will vary with the position along the aligned sequences' length (Claverie, 1992). It is therefore wise to consider several descriptors simultaneously. Likewise an HMM can be trained on both sequence and structure (Bystroff et al., 2000). HMMSTR models are based upon a library of protein sequence-structure motifs (Bystroff and Baker, 1998). Each state in an HMMSTR model emits an output symbol, representing sequence or structure. There are four categories of emission symbols: the traditional 20-letter amino acid alphabet, secondary protein structure (helix, strand or loop), 11 dihedral angle symbols and 10 structural context symbols.

This research explores structural DNA alignments (Chapters 8 and 9), encoded by structural alphabets that place an octamer's minimum energy conformer into a discrete bin. Flexibility is used to define inter-bin relationships, accounting for DNA dynamics. During this research, a related piece of work was found (Hasan, 2003), where DNA is translated into and aligned by flexibility sequences. Two shortcomings of this method should be pointed out. Only flexibility in terms of a tetranucleotide's slide (Packer et al., 2000b) is considered. What about actual conformations and other degrees of freedom? Secondly, inter-bin relationships are not defined, making sequence comparisons unrealistic.

7.7. Conclusions

The use of HMMs within biological sequence analysis is clearly a vast research area that has been under investigation and successfully used for well over a decade. Several bioinformatics applications exist, including protein family recognition, prediction of DNA-protein binding sites, gene location and identification of splicing sites. Bilmes points out the potential for HMMs to "accurately model any real-world probability distribution" (Bilmes, 2002). A major gap in the research area is pure structure based DNA HMMs.

When building structural HMMs many lessons can be learnt from previous sequence HMM research. The commonly used training algorithms (Baum-Welch, Simulated Annealing and Viterbi training) often have problems getting stuck in local optima due to a highly complex parameter space. This large number of free parameters can be partially controlled by the number of observables, model surgery and by designing simpler model architectures. Optimising by alignment quality or correct classification instead of the log odds ratio may be worth considering. Finally the quantity and quality of the training data is very important. Prior knowledge, in the form of substitution matrices, pseudocounts or Dirichlet mixtures, and sequence weights should be used to refine the quality. In deciding which descriptors to use, studying several simultaneously may be the most beneficial.

Chapter 8:
Structural DNA Alignments

The novel structural DNA alignment technique is introduced and its implementation discussed. The current methodology aligns sequences by a single minimum energy parameter. 3-step roll has been chosen as a starting point. Flexibility is encoded within a model's prior knowledge, in order to make comparisons between sequences and to consider the dynamic structure of DNA. Representation of the observable, the null hypotheses and prior knowledge via a substitution matrix are all topics of discussion. Methods for assessing the performance and predictive ability of models are presented. Finally a test scenario is used to ensure this novel extension to hidden Markov models is fully functional.

8.1. Structural Alphabets

In reality, the minimum energy parameters are continuous variables, however they are modelled here as being discrete, so as to reduce the algorithmic complexity and increase computational speed by avoiding the use of multivariate weighted Gaussian mixtures. In speech recognition, discrete versus continuous corresponds to speed versus accuracy, with accuracy and therefore continuous winning (Melnikoff et al., 2002). Both of these opposing factors are important in DNA analysis, since reliable results are desired from a tool that can digest the vast amount of biological information publicly available. The speed-accuracy dilemma does not always exist. It has been found that in some situations discrete representations outperform continuous in both accuracy and speed, for example in face recognition (Wallhoff et al., 2001) and in handwriting recognition (Rigoll et al., 1996). The logical approach taken here, due to the uncertainty of the best observable representation for structural DNA analysis, is to first assess the performance of the quicker discrete method. If the level of recognition is highly successful then the development of the much slower and perhaps no more accurate continuous approach will not be needed. Discrete probability distributions of the parameters will also allow a more direct comparison between this novel technique and

the traditional sequence alignment procedures that use a four-letter alphabet to represent the nucleotides. Structural alphabets have therefore been generated to represent the minimum energy structure of an octamer, with 3-step roll being solely considered as a simple starting point.

Discrete one degree values of 3-step roll have been previously calculated for the entire octamer population (Gardiner et al., 2003) and cover a range of –3 to +21°. This naturally forms a 25-letter alphabet of one degree bin widths, namely the A-Y alphabet. Note that the larger the alphabet size the larger the number of model parameters to estimate with more training data required for reliable models. Therefore the nucleotide alphabet will have an unfair advantage over the A-Y alphabet. For this reason a variety of structural alphabet sizes were created by grouping roll values into larger bins, allowing the effect of alphabet size upon model performance to be explored. Four different 3-step roll alphabets will be studied within this work (Table 8.1). Note that the 25 one degree bins can only be evenly grouped into five and that a structural alphabet of size four (A-D) has been included for closer analogy of structure to sequence DNA alignment.

Table 8.1: *The Structural 3-Step Roll Alphabets*

Alphabet Name	Size	Bin Width Description
A-D	4	1st 3 bin widths=6° and 4th=7°
A-E	5	All bin widths=5°
A-M	13	Bin widths=2° except last of 1°
A-Y	25	All bin widths=1°

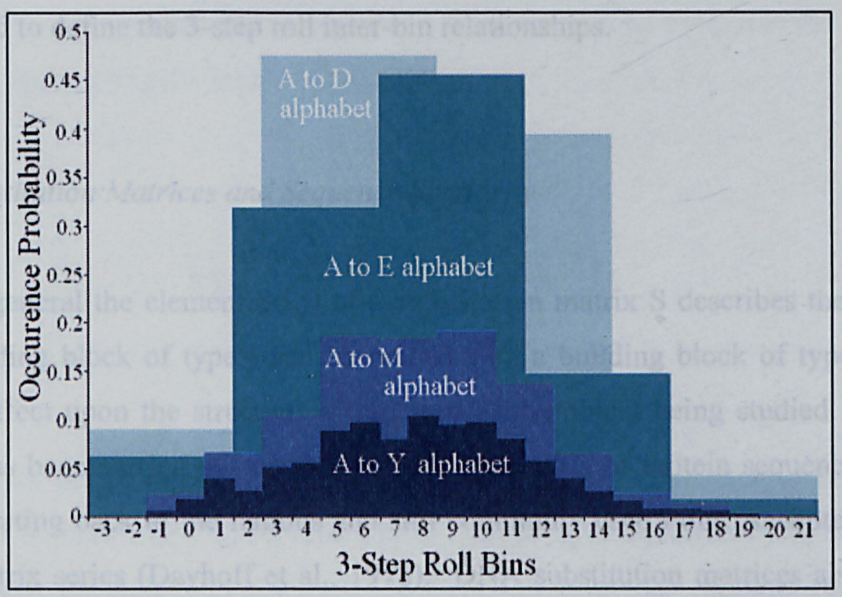
8.2. The Null Hypotheses

A null hypothesis is used to assess the meaningful connection between a sequence x and an HMM as opposed to a chance connection. It is the dominator of the odds ratio score. The sequence null hypothesis presented in Chapter 7 generates each nucleotide with a flat probability of ¼. If a structural hypothesis was based upon a similar flat distribution then the importance of the most commonly occurring roll bins would be exaggerated and vice versa for the rarer roll values. Instead the normalised

minimum energy 3-step roll frequencies must be used and will obviously vary with the alphabet size (Figure 8.1). The random generation of sequence x from the null model, *P(x|null)*, is given in Equation 8.1. *N* is the sequence length, x_i is the i^{th} value of roll in the sequence and f_{xi} is the frequency of the 3-step roll bin associated with x_i.

$$P(x \mid null) = \prod_{i=1}^{N} f_{xi} \qquad \textbf{Equ. 8.1}$$

Figure 8.1: *Frequency Distributions of The Structural Alphabets*



8.3. Inter-bin Relationships And Prior Knowledge

The HMM training procedure itself has no knowledge of the inter-bin relationships within each of the 3-step roll alphabets. It sees the placement of a 3-step roll from bin A and a 3-step roll from bin Y in an alignment column dominated by B's as equally favourable. The letters are either an identical match or a total mismatch. There is therefore a need to define the similarity between neighbouring bins as being higher than between distant ones. This is accomplished by adding prior knowledge into the HMM procedure, thereby suitably altering the state emission probabilities. Either Dirichlet priors (Sjolander et al., 1996) or substitution matrices (Dayhoff et al., 1978) address the analogous problem in protein sequence alignment, describing the similarity between amino acids with respect to their chemical features, size and shape.

Dirichlet priors have a major advantage over substitution matrices in protein analysis. They can represent the similarity between amino acids in several different contexts via their multi-component structure. One component might be concerned with hydrophobicity and another might be based upon similar sizes, reflecting the variable importance of several factors across an alignment. A substitution matrix has fixed amino acid similarities and therefore can only model similarity in a single context. However, when concerned with the structural alignments of this work a single component method will suffice, since only one context of similarity exists when measuring the distance between pairs of roll bins. For this reason a substitution matrix will be used to define the 3-step roll inter-bin relationships.

8.3.1. Substitution Matrices and Sequence Similarity

In general the element S(i,j) of a substitution matrix S describes the likelihood that a building block of type j can be replaced by a building block of type i with no dramatic effect upon the structure or function of the object being studied. Extensive research has been carried out upon substitution matrices of protein sequences for over 25 years, dating back to the famous and still commonly used Point Accepted Mutation (PAM) matrix series (Dayhoff et al., 1978). DNA substitution matrices also exist but have not received as much attention as those of proteins, probably due to their much smaller alphabet size. An example of a four by four DNA matrix that favours the alignment of purines and pyrimidines with themselves but not with each other is given in Table 8.2a (Lesk, 2003) along with a matrix that lacks any prior knowledge (Table 8.2b).

Table 8.2: *Examples of DNA substitution matrices*

a) Purine & Pyrimidine Matrix [Lesk 2003]

	A	G	T	C
A	20	10	5	5
G	10	20	5	5
T	5	5	20	10
C	5	5	10	20

b) Matrix lacking any prior knowledge

	A	G	T	C
A	1	0	0	0
G	0	1	0	0
T	0	0	1	0
C	0	0	0	1

A PAM matrix (Dayhoff et al., 1978) gives the probability that an amino acid will be replaced by another amino acid after a defined evolutionary time. Phylogenic trees were generated to study the evolutionary changes within 71 groups of closely related proteins and to determine the frequencies of the point accepted mutations. These frequencies were converted into mutation probabilities and then into the final log odds ratio PAM matrix that accounts for the random chance occurrence of the amino acid pair. Note that the matrix entries are commonly multiplied by a constant y and rounded to the nearest integer. The theory behind PAM construction forms the basis of subsequent protein matrices, which all have the general log odds ratio form given in Equation 8.2. *P(i,j)* is the probability that j will be replaced by i, f_i is the occurrence probability of i and *b* is the logarithmic base.

$$S(i,j) = y \log_b \left(\frac{P(i,j)}{f_i f_j} \right) = y \log_b \left(\frac{P(i \mid j)}{f_i} \right) \qquad \textbf{Equ. 8.2}$$

In this general form of a substitution matrix, a positive entry means that the substitution is more likely to be meaningful than to have occurred by random chance and vice versa for negative entries. It has been suggested that the matrices should be adjusted to the amino acid composition of the proteins being analysed instead of just using the standard background frequencies upon which they were originally built (Yu et al., 2003).

Well-known competitors to the PAM series are the Block substitution matrices (BLOSUM) (Henikoff and Henikoff, 1992). Like PAM, they are dependent upon the pre-alignment of proteins, but the sequences used are less similar to one another and are aligned in clustered blocks. The best matrix varies with the data being analysed, so a variety should be used with significance measured by a combined scoring scheme rather than the single highest matrix score (Frommlet et al., 2004). An alternative optimisation procedure is introduced that uses Bayesian decision theory to classify sequences, maximising the classification accuracy of a matrix (Hourai et al., 2004).

A dipeptide substitution matrix was constructed, in order to investigate any dependence of an amino acid substitution on neighbouring substitutions (Gonnet et al., 1994). PAM theory assumes there to be no such dependence, but is proven wrong. This extended matrix cannot however be applied to HMM analysis. A review of the different ways that amino acid similarities can be encoded has been carried out (Vogt et al., 1995) and includes volume comparisons, secondary structure comparisons and genetic code distances. Matrices based upon superimposed protein pairs have also been explored, where amino acids are said to be equivalent if C^{α} or C^{β} atoms are less than five Angstroms apart (Prlic et al., 2000).

Another example of incorporating structural information into a matrix is the linear combination of BLOSUM50 with a table of threading energies (Teodorescu et al., 2004). The threading energy of a protein is the energy it requires to form a shape analogous to another. An alternative use for substitution matrices is illustrated by the CLASSUM series that describe how an amino acid substitution within a particular protein family is involved in altering some functionality (Vilim et al., 2004).

8.3.2. Substitution Matrices within HMM analysis

A well known software package for producing HMMs is HMMER (Eddy, 1998). HMMER uses the "substitution matrix mixture" strategy (Durbin et al., 1998) to alter the emission probabilities with regard to the prior knowledge contained with a substitution matrix. This procedure is summarised by the 3 stages below.

(1) Convert all the substitution matrix entries, *S(i,j)*, to *P(i|j)*. Note that the matrix entries must first be converted to natural logarithms so that the exponential can then be taken. This is done by the scale factor ln *b*.

$$P(i \mid j) = f_i \exp\big(S(i,j) \times scale\big), \qquad \textbf{Equ. 8.3}$$

$$\text{where } scale = \frac{\ln b}{y}$$

(2) Calculate the pseudocount for letter i in alignment column a (α_{ia}), where f_{ja} is the fractional abundance of j in the column and A is the pseudocount weight with the HMMER default value of 20.

$$\alpha_{ia} = A\sum_{j} f_{ja} P(i \mid j) \qquad \textbf{Equ. 8.4}$$

(3) From the pseudocounts calculate the emission probability of i from the match state associated with column a, $e_a(i)$, where c_{ia} is the frequency count of i in the column.

$$e_a(i) = \frac{c_{ia} + \alpha_{ia}}{\sum_{i'} c_{i'a} + \alpha_{i'a}} \qquad \textbf{Equ. 8.5}$$

8.3.3. Substitution Matrices and Structural Similarity

The substitution matrix S depends on the occurrence frequencies (calculated for the null hypothesis), the two constants y and b (that shall be set to 3 and 2 respectively) and on one remaining variable yet to be determined, $P(i,j)$. This probability value (traditionally representing the mutation of one amino acid to another via an evolutionary process) represents the structural similarity between an octamer from bin i and an octamer from bin j. Two subtly different strategies for calculating $P(i,j)$ based upon DNA flexibility will be studied: the $P(i \leftarrow j)$ strategy and the $P(i=j)$ strategy.

- *The $P(i \leftarrow j)$ Strategy*:

  $P(i,j)$ is defined as the probability of an octamer in bin j flexing to a structure within bin i, $P(i \leftarrow j)$, analogous to the previously mentioned protein-threading energies.

- *The $P(i=j)$ Strategy:*

  $P(i,j)$ is defined as the probability that an octamer in bin i and an octamer in bin j will have the same structure, $P(i=j)$. $P(i=j)$ will be highly correlated with $P(i \leftarrow j)$, but it describes the dynamic structure of octamer pairs, giving a more realistic representation of DNA structural similarity than $P(i \leftarrow j)$.

Chapter 4 introduced a novel method for calculating the probability that one particular octamer will have a certain roll structure (section 4.1) or that two octamers will have the same roll structure (section 4.2). A highly generalised version of the same theory is used to calculate $P(i \leftarrow j)$ and $P(i=j)$, since a substitution matrix is not concerned with individual octamer probabilities, but with general tendencies of the octamers contained within particular roll bins.

First consider the matrices of the A-Y alphabet: $P_{AY}(i \leftarrow j)$ and $P_{AY}(i=j)$. Two major approximations are made. The average force constant over the entire octamer population (K_{av}, which equals 0.22 for 3-step roll) is used to represent the general flexibility of any octamer. Secondly, $P_{AY}(i \leftarrow j)$ uses a single rectangle of width one degree to approximate the required integral. The calculation therefore only involves a single exponential term (Equation 8.6). $P(i=j)$ is then derived from the $P(i \leftarrow j)$, see Equation 8.7. Note that flexibility is covered in the extended 3-step roll range of -20° to $+40^{\circ}$, as discussed in Chapter 4 section 4.1. There is a clear resemblance between Equation 8.7 and Equation 4.3.

$$P_{AY}(i \leftarrow j) = \frac{\exp\left[-K_{av}x^2 / RT\right]}{Q} \qquad \textbf{Equ. 8.6}$$

where $x=x_i-x_j$ (the distance between the bins being compared),

K_{av} is the average of $(3k^{-}_{roll} + 3k^{+}_{roll} / 2)$ for all octamers, 0.22

$$Q = \sqrt{2.5\pi / K_{av}}$$

$$P_{AY}(i = j) = \sum_{a=-20}^{40} P_{AY}(a \leftarrow i) P_{AY}(a \leftarrow j) \qquad \textbf{Equ. 8.7}$$

Once the two probability matrices above, $P_{AY}(i \leftarrow j)$ and $P_{AY}(i=j)$, had been calculated they were applied to Equation 8.2 with rounding to the nearest integer values, resulting in the substitution matrices $S_{AY,i \leftarrow j}$ and $S_{AY,i=j}$ respectively. The smaller alphabet matrices were then calculated by summing blocks in the P_{AY} matrices that refer to the larger roll bin widths. The block sums were then divided by either their widths

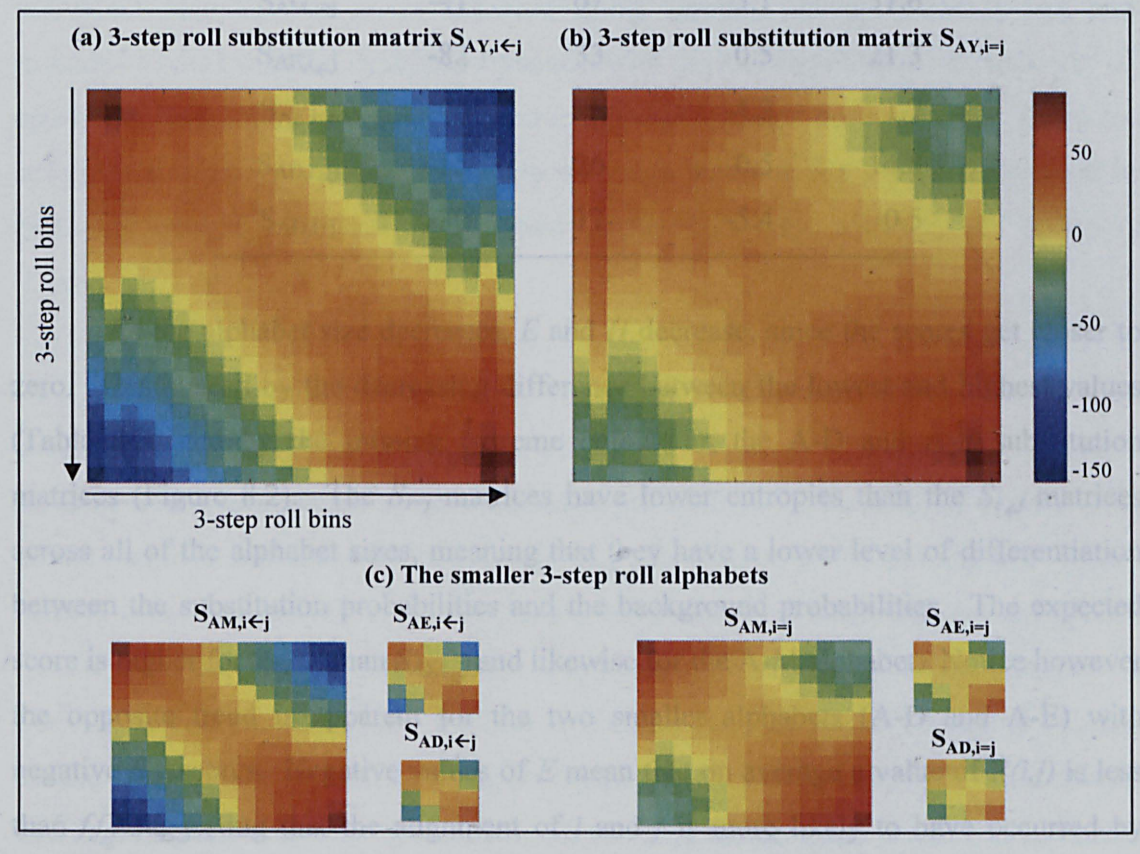
for $P(i \leftarrow j)$ or by their area for $P(i=j)$. This considers all combinations of ways that a larger bin can obtain a structure within another larger bin. For example, consider the calculation of $P_{AM}(A \leftarrow A)$ and $P_{AM}(A=A)$, shown in Equations 8.8 and 8.9.

$$P_{AM}(A \leftarrow A) = \frac{P_{AY}(A \leftarrow A) + P_{AY}(A \leftarrow B) + P_{AY}(B \leftarrow A) + P_{AY}(B \leftarrow B)}{2} \quad \textbf{Equ. 8.8}$$

$$P_{AM}(A = A) = \frac{P_{AY}(A = A) + P_{AY}(A = B) + P_{AY}(B = A) + P_{AY}(B = B)}{4} \quad \textbf{Equ. 8.9}$$

The resulting substitution matrices of the A-D, A-E, A-M and A-Y alphabets are shown in Figure 8.2. Note the anomalous lines in the S_{AY} matrices (figures 8.2a and b). These unusual features could be caused by the extremely low occurrence probabilities of bins B and X.

Figure 8.2: *3-step roll Substitution matrices. a) $S_{AY,i \leftarrow j}$ b) $S_{AY,i=j}$ and c) those of the smaller alphabets. Notice that values of S are all plotted on the same –150 to 75 scale so that direct comparisons can be made between all of the matrices.*



Two useful measures for describing the numbers within a substitution matrix are the expected score E (Henikoff and Henikoff, 1992) and the entropy H (Altschul, 1991). E is the average value within a matrix (Equation 8.10) and H measures the difference between the substitution probabilities and the background probabilities (Equation 8.11). E and H have both been calculated for all of the 3-step roll matrices (Table 8.3).

$$E = \sum_{i,j} \left(f_i f_j S(i,j) \right) \qquad \textbf{Equ. 8.10}$$

$$H = \sum_{i,j} \left(\frac{P(i,j) S(i,j)}{y} \right) \qquad \textbf{Equ. 8.11}$$

Table 8.3: *Lowest and highest values, expected scores (E) and entropies (H) of the matrices*

Matrix	Lowest	Highest	E	H
$S_{AY,i\leftarrow j}$	-151	82	4.8	201.9
$S_{AY,i=j}$	-43	81	8.9	180.9
$S_{AM,i\leftarrow j}$	-137	68	2.4	90.2
$S_{AM,i=j}$	-41	67	3.1	37.6
$S_{AE,i\leftarrow j}$	-82	33	0.5	21.3
$S_{AE,i=j}$	-40	25	-4.2	1.7
$S_{AD,i\leftarrow j}$	-55	26	0.5	14.8
$S_{AD,i=j}$	-30	17	-5.4	0.5

As the alphabet size decreases, E and H decrease, since the scores get closer to zero. This is seen by the decreasing difference between the lowest and highest values (Table 8.3) and by the missing extreme colours in the A-D and A-E substitution matrices (Figure 8.2). The $S_{i=j}$ matrices have lower entropies than the $S_{i\leftarrow j}$ matrices across all of the alphabet sizes, meaning that they have a lower level of differentiation between the substitution probabilities and the background probabilities. The expected score is higher for $S_{AY,i=j}$ than $S_{AY,i\leftarrow j}$ and likewise for the A-M alphabet. Notice however the opposite trend is apparent for the two smaller alphabets (A-D and A-E) with negative $S_{i=j}$ scores. Negative values of E mean that on average a value of $P(i,j)$ is less than $f_i f_j$, suggesting that the alignment of i and j is more likely to have occurred by chance than because of structural equivalence. $S_{AY,i=j}$ sees the possibility that distant roll

bins can be structurally equivalent, whereas $S_{AY,i \swarrow j}$ strongly forbids distant bins to have equivalent structures (note the presence of dark blue in the corners of the matrix). An alternative approach to calculating a 3-step roll substitution matrix is a simple linear probability scale based upon differences in roll (the probability being inversely proportional to $|x_i - x_j|$). This could be thought of as describing a protein's view of the DNA sequences, which could be useful when analysing a particular set of protein binding sequences.

8.4. Software

HMMER version 1.8.4 (Eddy, 1998) is commonly used to generate traditional sequence HMMs and alignments. Here, it has been extended and generalised to deal with the 3-step roll alphabets, their null hypotheses and substitution matrices. Four executable programs from the HMMER package are used throughout this work: 'hmmt', 'hmma', 'hmme' and 'hmms'. The 'hmmt' program generates trained models, 'hmma' then produces sequence alignments from the models and 'hmme –b' emits the model's most probable sequence, also referred to as the consensus sequence. Finally, individual sequence scores are obtained using 'hmms'. Here, HMMER has been extended to deal with the structural alphabets with the null hypotheses encoded. A '-Z' option has been implemented for structural alphabet selection purposes. The 3-step roll substitution matrices have each been placed into a common file format understood by HMMER version 1.8.4 and are specified in the usual way, with the '-P' option of 'hmmt'.

Default parameters were used to construct all models and are as follows. The traditional HMM architecture discussed in Chapter 7 section 7.3 is used with unbiased uniform state transition and symbol emission probabilities in the starting models. The model length is initialised to the average sequence length of the training set. Model surgery is used to optimise the model length and a simulated annealing strategy (SA) is used to train the model whilst attempting to avoid local minima problems (Chapter 7). The SA default parameters are a kT of five and a ramp of 0.95. The value of ramp defines the 'cooling' process by being the factor that kT decreases by upon each training iteration. Finally once SA has converged, the Viterbi algorithm is used to refine the alignment solution.

8.5. Evaluating the matrices

Before discussing the methods used to assess model performance and reliability, an evaluation of the 3-step roll substitution matrices (section 8.3.3) is presented. The alignment of two sequences from the A-Y alphabet is shown below, with each correctly aligned column composed of two octamers from adjacent roll bins. Note that a full stop symbol refers to a gap.

```
..ADJNQRMT..
SSBEKMPSNSAA
```

This pair of sequences has been chosen to check that the substitution matrices recognise the inter-bin relationships that are essential for their correct alignment. Alignments are also obtained using no prior knowledge, a substitution matrix of random probabilities, a randomly shuffled $S_{AY,i\leftarrow j}$ matrix and a matrix with probabilities linear to the roll bin distances (where $P=constant/|x_i-x_j|$). The linear matrix was mentioned previously in section 8.3.3 as describing a protein's view of DNA sequences. The first sequence (ADJNQRMT) was hard-wired into the training set by duplicating it 19 times. 100 models were then generated for each prior knowledge type with the frequency of the correct alignment counted and its score noted (Table 8.4).

Table 8.4: *Assessing varying levels of prior knowledge for the A-Y alphabet and the alignment of two sequences, ADJNQRMT and SSBEKMPSNSAA. Note that lowercase letters come from an insert state.*

Prior Knowledge	Alignment	Score	Frequency
None	ADJ....NQRMT SSBekmpSNSAA	36.07	ABSENT
$S_{AY,i\leftarrow j}$	..ADJNQRMT.. SSBEKMPSNSAA	37.76	87%
$S_{AY,i=j}$	..ADJNQRMT.. SSBEKMPSNSAA	37.12	32%
Randomised $S_{AY,i\leftarrow j}$	A......DJNQRMT .ssbekmPSN.SAA	35.03	ABSENT
Random probabilities	ADJ....NQRMT SSBekmpSNSAA	36.47	ABSENT
Linear with roll bin distances	ADJ....NQRMT SSBekmpSNSAA	36.26	ABSENT

The two substitution matrices based on structural DNA probabilities ($S_{AY,i_{\leftarrow}j}$ and $S_{AY,i=j}$) are the only two methods that align the test sequences correctly (Table 8.4). $S_{AY,i_{\leftarrow}j}$ is more reliable than $S_{AY,i=j}$ giving the correct alignment 87% of the time in comparison to 32%. However, the best performing substitution matrix may vary with the sequences considered. Surprisingly, the matrix based on probabilities that are linear with the roll bin distances results in the same highest scoring solution as the random matrix or as no prior knowledge at all in this instance.

8.6. Model Assessment

It is important to be able to measure the performance and robustness of models generated, in order to judge whether suggested alignments are reliable. Such assessment techniques will also prove useful for comparing sequence alignments to structural alignments. Three methods have been chosen: the non-validated approach, leave-one-out cross validation (LOO CV) and test set validation.

8.6.1. The non-validated approach

This approach uses the entire dataset as a training set and analyses the distribution of model scores obtained from 100 HMMs. A model score is the average log odds ratio of all the sequences in the dataset (Chapter 7) and is therefore a measure of how well the training data is explained by a model. The higher a score, the better the data is explained. The standard deviation of the score distribution over 100 models measures the reproducibility and precision of an alignment solution. The distribution's mean is the average performance and strength of explanation across the models. The best scoring model is used to generate the final alignment solution, which is viewed by a matrix plot and summarised by a logo plot (Chapter 6).

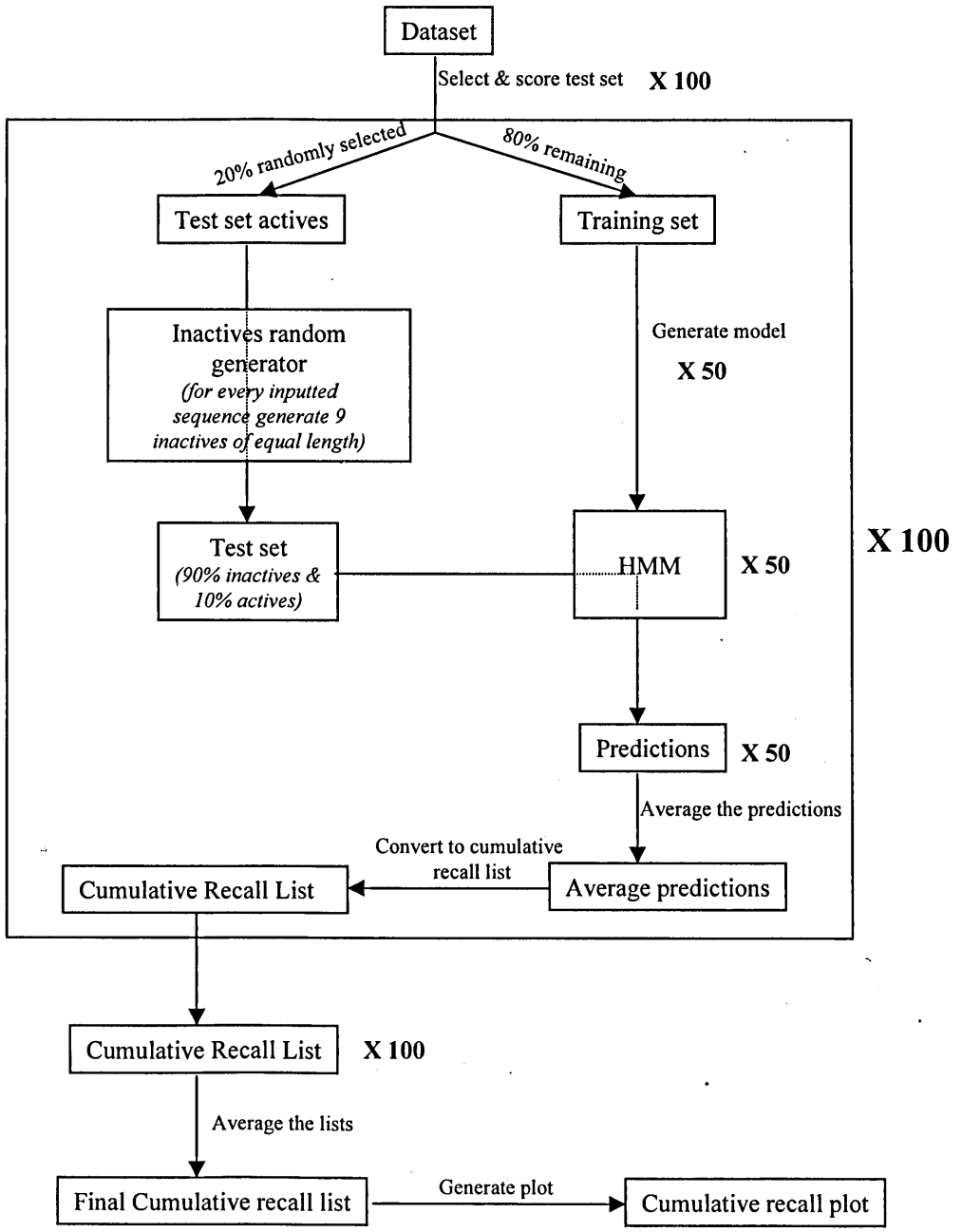
8.6.2. Leave-one-out Cross validation (LOO CV)

Not only is it important to have a model that explains the training data well (as measured by the above non-validated analysis). An estimation of how well it might fit data outside of the training set (its predictive ability) is also needed. For this purpose leave-one-out cross validation (LOO CV) is performed. Every sequence within a dataset is removed in turn and its score predicted by 100 HMMs generated in its absence. The average of these 100 values is then taken as a single cross validated sequence score (the score for sequence x being $S_{cv,x}$). The values $S_{cv,1}$ to $S_{cv,n}$ (n being the dataset size) are then averaged to give an overall LOO CV score for the whole dataset. The amount of decrease from the non-validated model scores to the LOO CV model scores corresponds to robustness. If a model is very robust then removing one sequence will not largely affect its predictive ability, the decrease being small. However if the dataset contains a lot of outliers then removing a sequence from the training data will lead to poor predictions and a much smaller LOO CV model score with a high likelihood of model overfitting (Hawkins, 2004).

8.6.3. Test set validation

This final data analysis technique measures the performance of an HMM by applying it to an external test set that is composed of both active and inactive sequences. 20% of the dataset is randomly selected as the test set actives with the remaining 80% forming the training set. For each selected active, nine random presumed inactive sequences of equal length are generated, finally resulting in a test set of 90% inactives and 10% actives. 50 models are generated and each used to score the test set sequences, since the same training data can lead to different HMM solutions. The results will also depend upon the test set used, so in order to remove any bias 100 test sets were randomly generated. The scores from the resulting 5000 models per dataset were combined in the following manner. Firstly, for each particular test set run the 50 scores obtained for each test set sequence were averaged. The test set sequences were then ordered by their decreasing average scores and converted into a ranked cumulative recall list. The 100 ranked cumulative recall lists can then be averaged and a single cumulative recall plot finally generated. This technique is summarised in Figure 8.3.

Figure 8.3: *The test set validation procedure. 100 test sets are randomly selected and analysed. Their recall results are then combined into a final cumulative recall list from which a plot is generated. The large central box contains the analysis procedure carried out individually on each of the 100 test sets and involves the generation of 50 HMMs per test set.*



There has been much discussion of LOO CV versus test set validation (Shao, 1993; Hawkins et al., 2001; Hawkins et al., 2003; Hawkins, 2004). Evidence has been given that when a dataset is small removing a subset as an external test set is not sensible, since essential information will be lost from the training data (Hawkins et al., 2003). The results may also be unreliable due to their high variation with the test set selected. However, since the procedure used here involves selecting 100 test sets randomly it can be likened to a k-fold cross validation. The only difference being that rather than choosing the test sets by splitting the dataset into k-partitions they are chosen independently from one another. It has been argued that LOO CV tends to overestimate predictivity and that k-fold cross validation gives a more reliable estimation (Shao, 1993).

The measure $Recall_{NORM}$ (Salton and McGill, 1983) can be used to assess a model's recall ability in comparison to the perfect scenario (Equation 8.12). A value of one represents perfect recall and zero is the worse case scenario (all actives being placed at the bottom of the recall list).

$$Recall_{NORM} = 1 - \left[\frac{\sum_{i=1}^{ACT} RANK_i - \sum_{i=1}^{ACT} i}{ACT(N - ACT)} \right] \qquad \textbf{Equ. 8.12}$$

where ACT is the number of active sequences in the test set, $RANK_i$ is the rank of i^{th} active sequence and N is the test set size.

Three different methods can be used to generate the nine inactive sequences associated with each active. The first randomly generates sequences having the same length as the active, selecting one of the four nucleotides at each position with an independent probability of 0.25. The second and third methods preserve the nucleotide and dinucleotide composition respectively by performing hundreds of suitable shuffling operations on the active sequence. The operation involved in retaining the mononucleotide frequency distribution is simply to randomly select and swap the positions of two nucleotides. Dinucleotide shuffling however is more complicated, because swapping random pairs of doublets will alter the overlapping composition

within the sequence. A simple solution to this was used, the swapping of two sub-sequences that have identical starting letters and ending letters. For example in the sequence GG**A**CAT**G**GTTAT**A**ATTT**G**CTAG, the two highlighted sections can be swapped since they both start with A and end with G. It was ensured that each inactive generated was not the same as the original sequence or the other associated inactives. Other more sophisticated techniques include an algorithm that uses a Markov chain structure in its implementation (Kandel et al., 1996) and a method involving Eulerian walks and graph theory (Altschul and Erickson, 1985; Coward, 1999; Wu and Gu, 2002).

Recall is related only to the score ordering and lacks details of the score values and therefore the degree of separation between actives and inactives. However, the difference between the active score distribution and inactive score distribution can be measured and assessed by T-tests (Chapter 5). Unpaired heteroscedastic two-tailed T-tests (Miller and Miller, 1994) are therefore performed on the active and inactive score distributions obtained in test set analysis. The null hypothesis of two population means being equal is used and a confidence level of 95%.

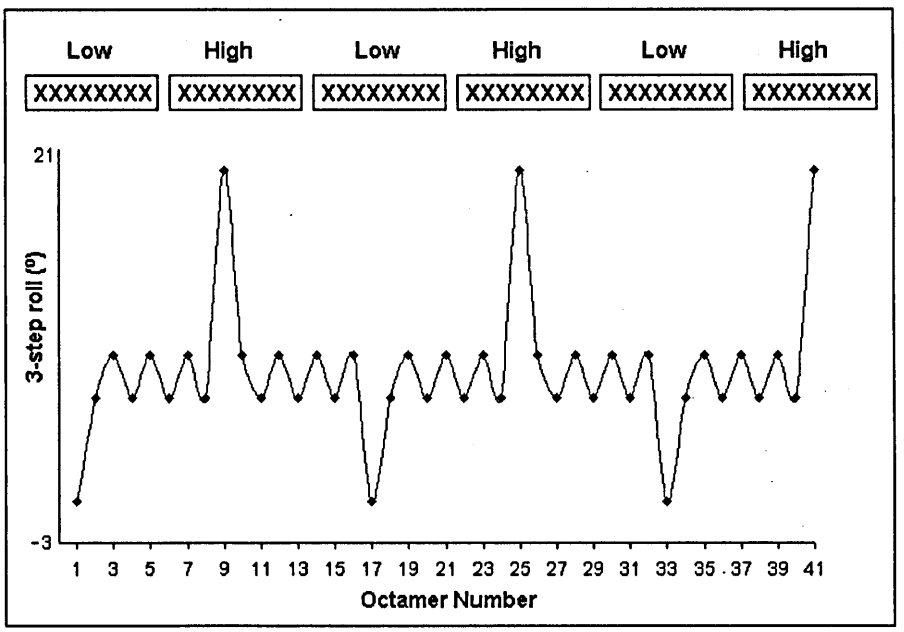
8.7. Artificial Dataset

The functionality of the novel structural DNA alignment technique must be fully tested before applying it to real data. Therefore an artificial dataset is created here, in order to set up a test scenario with a known correct outcome. The use of the three model assessment techniques introduced in section 8.6 will be illustrated. The artificial dataset has been designed so that it is more conserved by a 3-step roll motif than by its nucleotides. Therefore it is expected that for this mock run the structural alignment results should be superior to traditional sequence analysis. This will act as a null test. If the outcome is unsuccessful here then the structural HMM methodology is incorrect and will need to be refined.

8.7.1. Creating the dataset.

Creating a dataset that is more conserved by 3-step roll than by sequence requires vertical blocks of octamers to be very similar by roll but different by their nucleotides. Note that there are seven intervening octamers determined by the two adjacent non-overlapping octamer columns (Figure 8.4). A decision was made to design a motif of alternating low and high 3-step roll. "Low" octamers were defined as those having a 3-step roll less than 3^{o} and "high" were defined as those with a 3-step roll greater than 14^{o}.

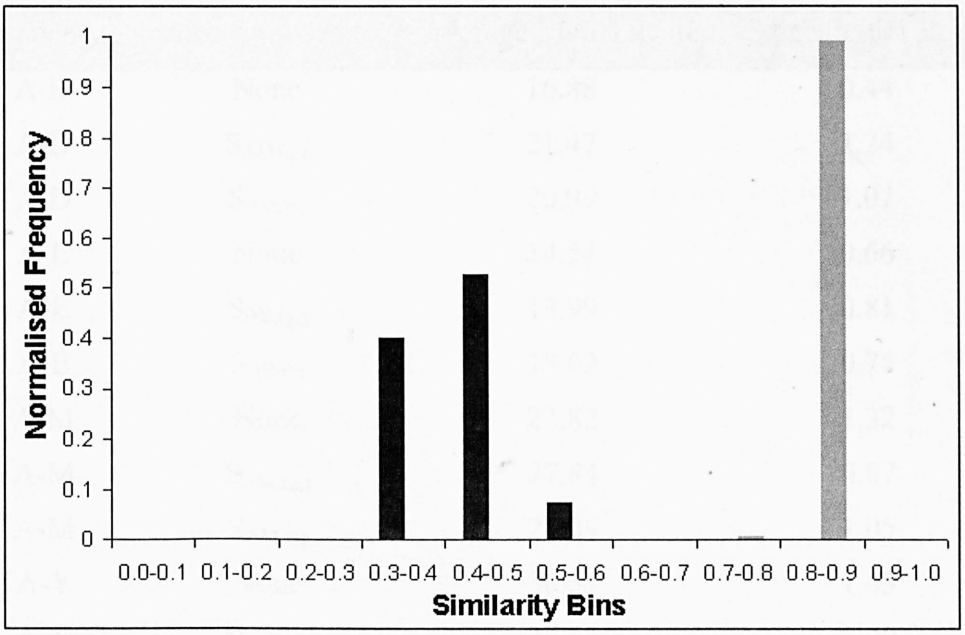
Figure 8.4: *Designing the artificial dataset with a motif of alternating low and high 3-step roll. There are seven intervening octamers pre-determined by each low and high octamer.*



Two diverse selections by sequence were made, one from the collection of octamers with "low" roll and the other from the octamers of "high" roll. The sphere exclusion clustering technique (Butina, 1999) was used, where all octamers within a specified sphere radius from a selected octamer are excluded. The distance between two octamers was defined by their sequence dissimilarity (the fraction of their nucleotides that mismatch). A sphere radius of 0.5 gave 48 "low" octamers and 49 "high" octamers, resulting in an artificial dataset of 16 "Low High Low High Low High" sequences of length 48 nucleotides.

Before constructing any models, a check was made that the dataset's sequence similarity was considerably lower than its structural similarity. The sequence similarity was calculated by averaging the pairwise Needleman-Wunsch similarity (Needleman and Wuncsh, 1970) of all possible sequence pairs within the dataset. The structural similarity was based on the 3-step roll distances between octamers and used an algorithm analogous to Needleman-Wunsch, in order to align pairs of sequences by maximising their pairwise structural similarities. The average sequence similarity of the artificial dataset is 0.417 and the average structural similarity is 0.841, therefore confirming that the artificial dataset is more conserved by structure (3-step roll) than by sequence. Note also that the distribution of pairwise similarities for sequence has a larger variance than structure (Figure 8.5). Application of this dataset to the novel structural alignment technique and to the traditional sequence alignment technique can now be carried out with confidence that the superior pattern recognition results should come from structure not sequence.

Figure 8.5: *Pairwise similarity distributions for sequence and structure, confirming that the artificial dataset is more conserved by 3-step roll (grey) than sequence (black). Needleman-Wunsch similarities are used for sequence and structure with the pairwise structural alignments based on minimising 3-step roll distances.*



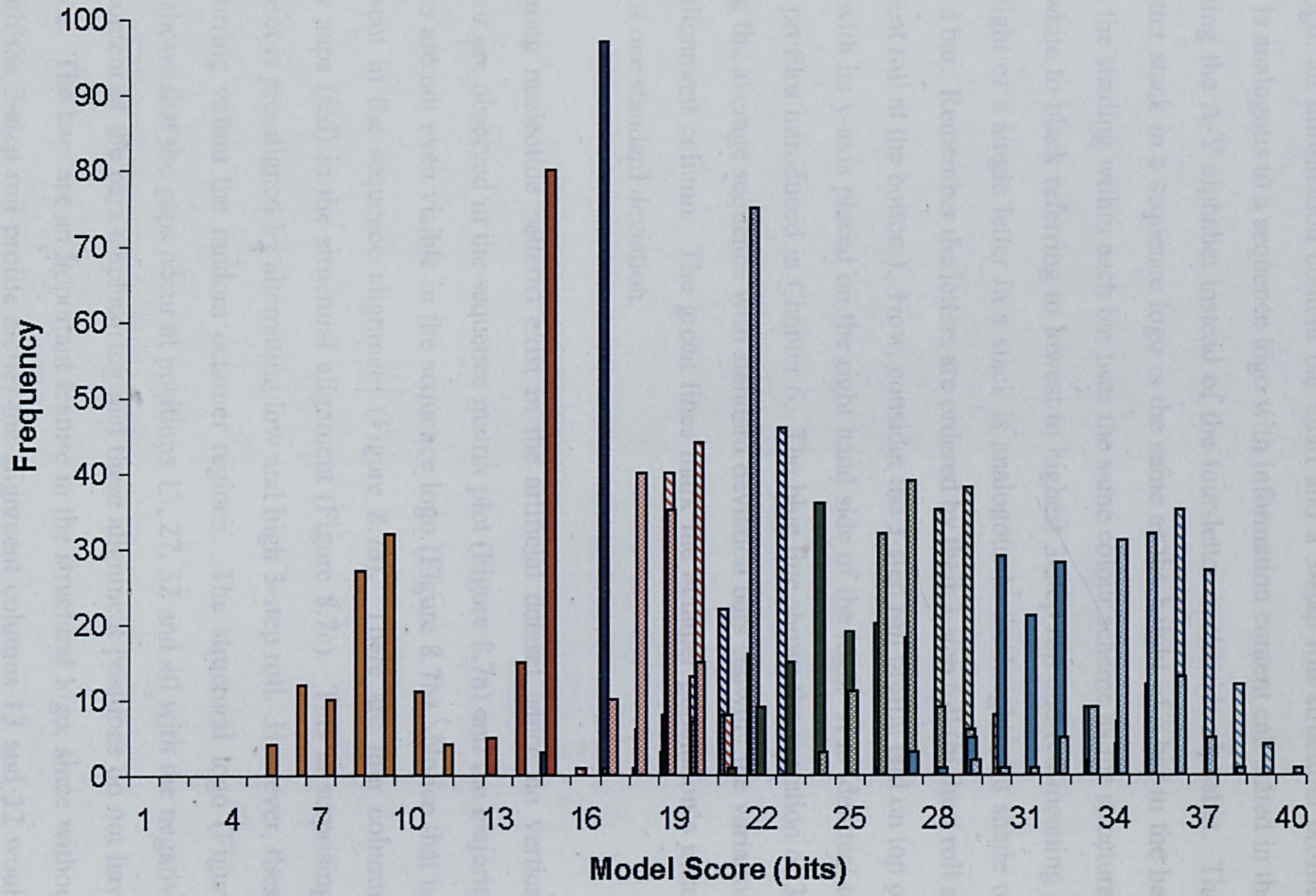
8.7.2. The non-validated approach

100 non-validated models were generated and the model score distributions were analysed for sequence and each of the four 3-step roll alphabets (A-D, A-E, A-M and A-Y) in combination with three types of prior knowledge (none, $S_{i\leftarrow j}$ and $S_{i=j}$). The average model score for sequence alignment is lower than any of the structural alignments, regardless of the level of prior knowledge used (Table 8.5 and Figure 8.6). The model scores tend to increase as the structural alphabet size increases, with the exception of A-D fitting the data slightly better than A-E. Within each alphabet a general trend is observed (Figure 8.6). The $S_{i\leftarrow j}$ substitution matrix is clearly performing better than the $S_{i=j}$ substitution matrix which in turn is performing better than no prior knowledge at all. The most precise model scores and reproducible models are generated with no prior knowledge and the two smallest alphabets (A-D having a standard deviation of 0.44 and A-E having a standard deviation of 0.66). This could be due to the larger alphabets having a larger number of model parameters to minimise which means that when inter-bin relationships are defined a variety of alternative favourable matches can be found (leading to less precise models).

Table 8.5: *Non-validated scores for artificial dataset with varying alphabet type and prior knowledge.*

Alphabet	Prior Knowledge	Average Model Score	Standard deviation
A-D	None	16.88	0.44
A-D	$S_{AD,i\leftarrow j}$	21.47	1.24
A-D	$S_{AD,i=j}$	20.99	1.07
A-E	None	14.54	0.66
A-E	$S_{AE,i\leftarrow j}$	18.99	0.81
A-E	$S_{AE,i=j}$	18.02	0.75
A-M	None	23.82	1.32
A-M	$S_{AM,i\leftarrow j}$	27.84	0.87
A-M	$S_{AM,i=j}$	26.09	1.05
A-Y	None	30.57	1.45
A-Y	$S_{AY,i\leftarrow j}$	35.84	1.29
A-Y	$S_{AY,i=j}$	33.90	1.54
Sequence	None	7.75	1.40

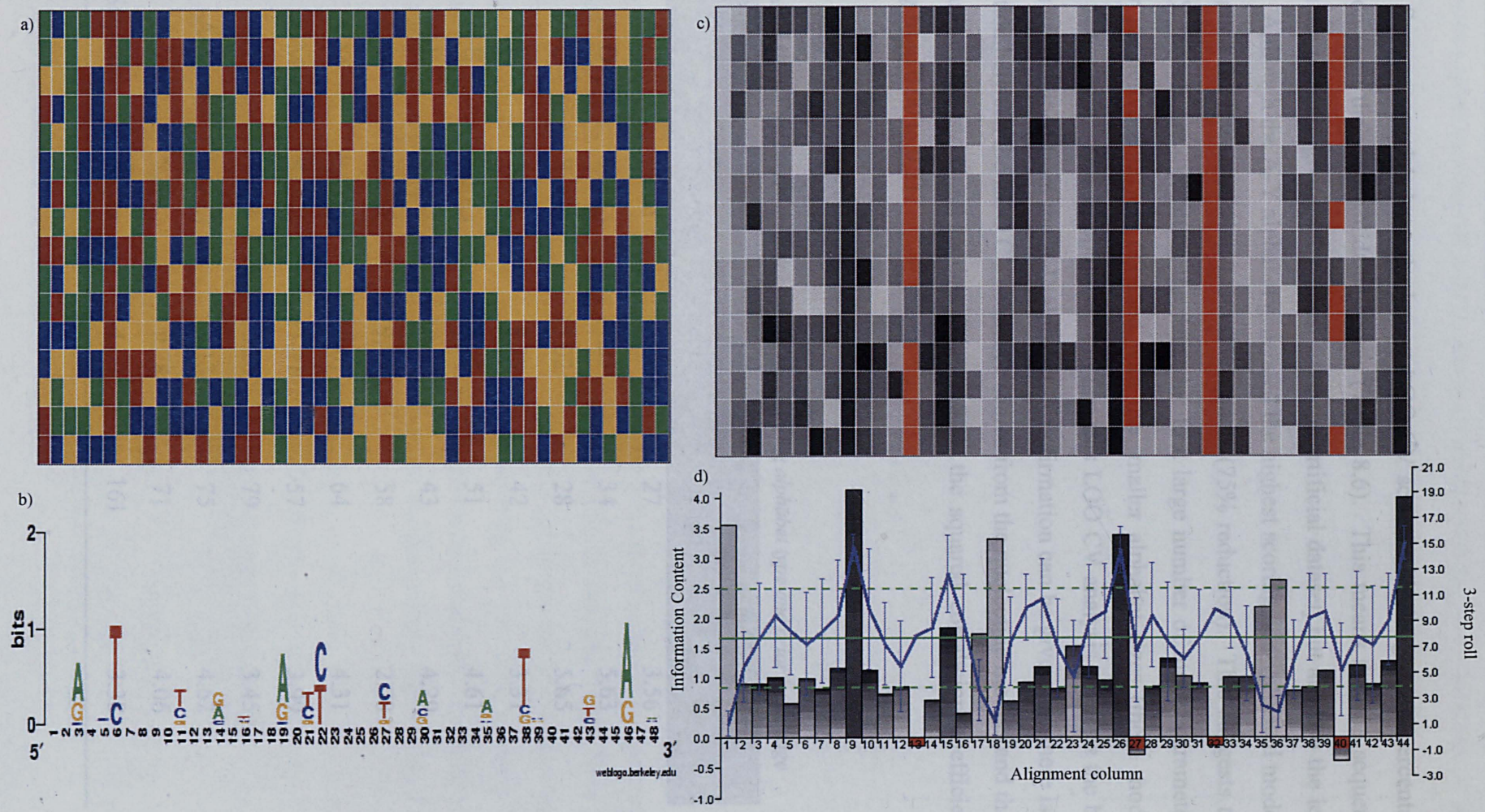
Figure 8.6: *Score distributions for 100 non-validated HMM runs across the different alphabets with differing levels of prior knowledge. The different alphabets are sequence (orange), A-D (dark blue), A-E (red), A-M (green) and A-Y (light blue). The type of bar shading refers to the different levels of prior knowledge: no prior knowledge (solid shading), $S_{i=j}$ substitution matrix (medium shading) and $S_{i \leftrightarrow j}$ substitution matrix (light shading).*



The highest scoring sequence model and 3-step roll model were selected and their alignment solutions were examined (Figure 8.7). A detailed explanation of matrix plots and sequence logos was given in Chapter 6. However, a novel structural logo is now introduced in Figure 8.7d and needs to be explained. There are two components to a structural logo: an information content bar chart and a 3-step roll structural profile. The bar chart is analogous to a sequence logo with information content calculated in the same way, using the A-Y alphabet instead of the four-letter nucleotide alphabet. The height of a letter stack in a sequence logo is the same as the height of a bar in the bar chart. Notice the shading within each bar uses the same colour scheme as the structural matrix plot (white to black referring to lowest to highest 3-step roll and red meaning a gap). The height of a single letter in a stack is analogous to the height of a shade of colour within a bar. Remember the letters are ordered by their 3-step roll (highest roll at the top to lowest roll at the bottom). Now, consider the 3-step roll profile laid on top of the bar chart with its y-axis placed on the right hand side of the logo. It is identical to the structural profiles introduced in Chapter 6. The blue line shows the variation of 3-step roll along the average sequence with standard deviation bars showing the variation within each alignment column. The green lines mark the octamer population's mean plus and minus one standard deviation.

No strong nucleotide patterns exist in the artificial dataset, since no vertical bands of colour are observed in the sequence matrix plot (Figure 8.7a) and the majority of letter stacks are not even visible in the sequence logo (Figure 8.7b). Notice that no gaps are present in the sequence alignment (Figure 8.7a). There are four columns dominated by gaps (red) in the structural alignment (Figure 8.7c). This is surprising, since the dataset is pre-aligned by alternating low and high 3-step roll. However, these gaps are occurring within the random octamer regions. The structural logo (Figure 8.7d) clearly shows that the gaps occur at positions 13, 27, 32 and 40 with the negative information content of the bars emphasising that these alignment positions do not have any importance. The bars are an important feature in the structural logo, since without them (and with the 3-step roll profile alone) the alignment columns 13 and 32 would appear to have perfect conservation with standard deviations of zero. The original 3-step roll motif upon which the artificial dataset was built can be seen from the high bars (positions 1, 9, 18, 26, 36 and 44) and their alternate light and dark shading.

Figure 8.7: *Sequence alignment versus structural alignment of artificial dataset. a)Sequence matrix plot, where blue = C, orange = G, green = A and red = T. b) Sequence logo, obtained from Weblogo (Crooks et al., 2004). c) Structural matrix plot, where red = gap and light to dark shading is low to high 3-step roll. d) Structural logo*



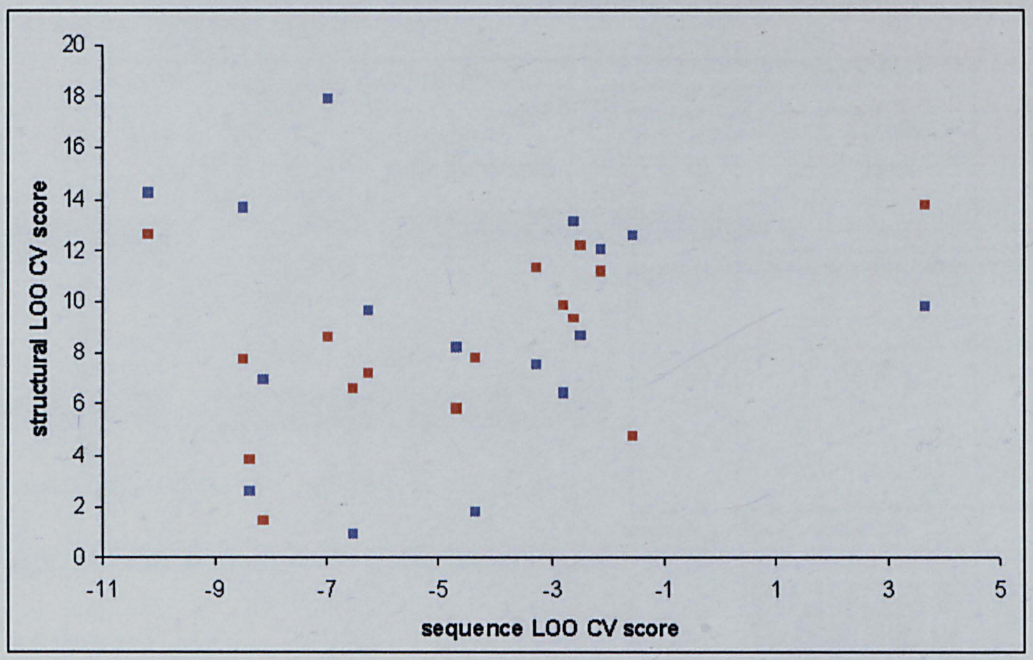
8.7.3. Leave-one-out cross validation

*S*equence models have the lowest LOO CV scores with the greatest percentage reduction from their non-validated scores (Table 8.6). This means that the sequence HMMs are not only giving the poorest fit to the artificial dataset, but are also the least robust. Although the A-Y alphabet generated the highest scoring non-validated models, it generated the lowest scoring LOO CV models (75% reduction). This suggests that the A-Y models are overfitting the data, due to a large number of model parameters. Greater predictive ability is obtained with the smaller alphabet sizes (those models having fewer parameters to optimise). In fact this LOO CV analysis suggests the best models are obtained from the A-D alphabet. Confirmation can be given that there is no correlation between the LOO CV sequence scores from the nucleotide models and those from any of the structural models (Figure 8.8), the squared correlation coefficients varying from 0.00 to 0.21.

Table 8.6: *LOO CV scores for artificial dataset with varying alphabet type and prior knowledge*

Alphabet	Prior Knowledge	LOO CV Score	% Score Reduction	Standard deviation
A-D	None	12.29	27	3.56
A-D	$S_{AD,i \leftarrow j}$	14.23	34	5.63
A-D	$S_{AD,i=j}$	15.02	28	5.65
A-E	None	8.39	42	3.31
A-E	$S_{AE,i \leftarrow j}$	9.38	51	4.61
A-E	$S_{AE,i=j}$	10.23	43	4.29
A-M	None	9.95	58	2.78
A-M	$S_{AM,i \leftarrow j}$	10.01	64	4.31
A-M	$S_{AM,i=j}$	11.21	57	3.99
A-Y	None	6.28	79	3.45
A-Y	$S_{AY,i \leftarrow j}$	9.13	75	4.62
A-Y	$S_{AY,i=j}$	9.78	71	4.06
Sequence	None	-4.69	161	3.38

Figure 8.8: *Correlation between LOO CV sequence scores and LOO CV structural scores. Red = A-E alphabet with no prior knowledge (R^2 is 0.21). Blue = A-Y alphabet with $S_{AY,i<j}$ prior knowledge (R^2 is 0.00).*



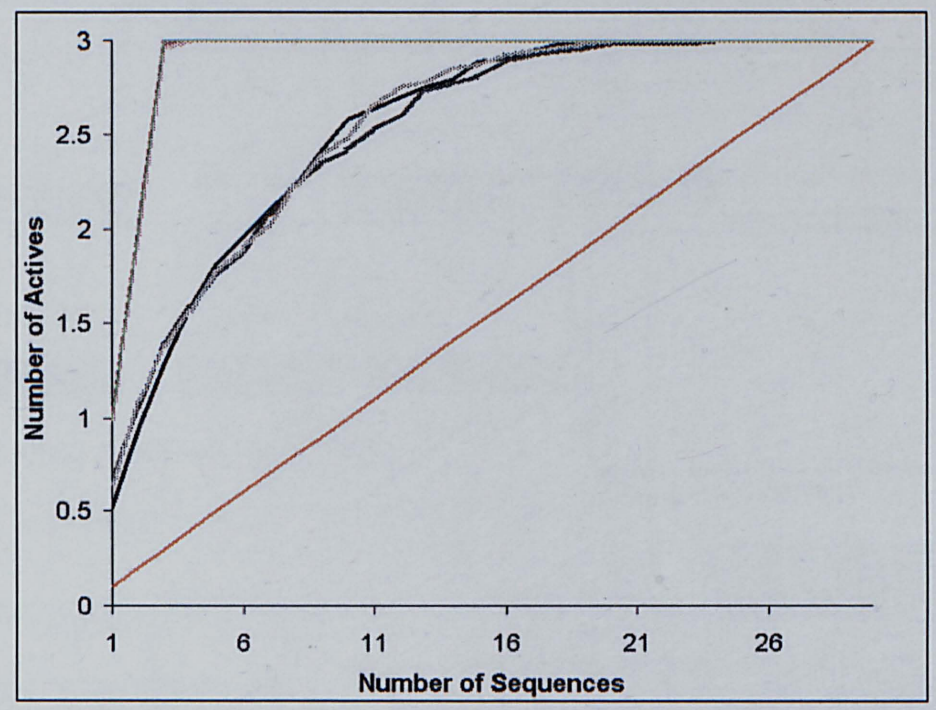
8.7.4. Test set analysis

Test set analyses were carried out as described in section 8.6.3 for all the alphabets (sequence, A-D, A-E, A-M and A-Y) using the three methods for generating the inactives (random, mono-nucleotide shuffling and dinucleotide shuffling). Note that the active sequence set is in fact the artificial dataset (or in general the dataset being modelled). The $S_{i<j}$ matrix was chosen when generating the structural models. All structural models had perfect recall ability (Figure 8.9) with $Recall_{NORM}$ equalling 1.00. The $Recall_{NORM}$ values for sequence were not perfect, but surprisingly high (Table 8.8).

Table 8.8: *Values of $Recall_{NORM}$ for sequence. N.B. (All values of $Recall_{NORM}$ for structure were 1.00*

Settings	Average $Recall_{NORM}$
Sequence with random inactives	0.864
Sequence with mono shuffled inactives	0.859
Sequence with doublet shuffled inactives	0.867

Figure 8.9: *Cumulative recall plot for the artificial dataset, considering the sequence models (black) and the A-Y roll alphabet models (purple). The $S_{i \leftarrow j}$ substitution matrix is used for structure and the three classes of inactives: random inactives (solid lines), mono shuffled inactives (medium weighted lines) and doublet shuffled (light weighted lines).*



Score distributions (Figure 8.10) of the active and inactive sequences show that structure clearly separates the actives from the inactives with positive and negative scores respectively, regardless of the method used to generate the inactives. Sequence models tend to score both actives and inactives as negative, although the distributions have a significant mean separation of around 5 bits with values of T greater than 20 (Table 8.9). The mean separation of the distributions from the structural models is around 35 bits with values of T geater than 105. Structure clearly differentiates the actives from the inactives to a much greater extent than sequence (the mean separation being about 7 times greater).

Figure 8.10: *Score distributions of the sequence inactives (red), sequence actives (green), structure inactives (purple) and structure actives (light blue). The level of shading refers to the method used to generate the inactives: random generation (solid lines), mono shuffling (lighter shading) and doublet shuffling (dashed lines). The $S_{i \leftarrow j}$ substitution matrix is used for structure.*

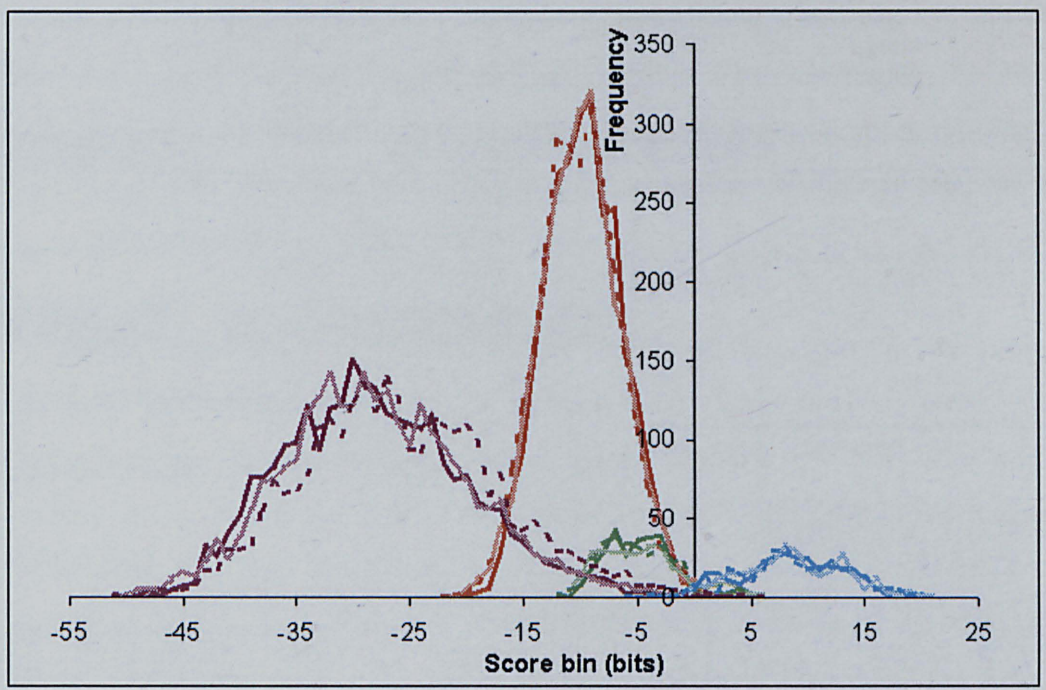


Table 8.10: *T-test analyses between active and inactive score distributions from Figure 8.10. DOF is the number of degrees of freedom. $\bar{x}_1 - \bar{x}_2$ is the mean separation between the two distributions. The $S_{i \leftarrow j}$ substitution matrix is used for structure.*

Alphabet	Type of inactives	DOF	T	$\bar{x}_1 - \bar{x}_2$
Sequence	Random	383	25.75	4.95
Sequence	Mono	362	23.67	5.33
Sequence	Doublet	383	26.08	5.09
A-Y	Random	539	119.53	36.57
A-Y	Mono	535	116.09	36.42
A-Y	Doublet	543	105.74	34.30

8.8. Conclusions

A novel HMM technique that successfully aligns sequences by their 3-step roll has now been introduced. Structural alphabets of different sizes (A-D, A-E, A-M and A-Y) form discrete representations of 3-step roll. Relationships between the bins in an alphabet have been encoded via substitution matrices that incorporate the general flexibility of DNA octamers into an HMM. $S_{AY,i\leftarrow j}$ was found to be more reliable than $S_{AY,i=j}$. The correct alignment of a sequence pair was obtained 87% of the time with $S_{AY,i\leftarrow j}$ in comparison to 32% of the time with $S_{AY,i=j}$.

Model performance will depend on the number of model parameters (alphabet size) and the amount and quality of the training data. Three methods were used to assess performance (the non-validated approach, leave-one-out cross validation and test set validation). An artificial dataset was applied to the novel alignment procedure, confirming that it is fully functional. In the non-validated approach, sequence scored the worst and A-Y the best. The most precise scores were, however, obtained using the A-D or A-E 3-step roll alphabet. The substitution matrix of choice is clearly $S_{AY,i\leftarrow j}$ rather than $S_{AY,i=j}$. No strong nucleotide patterns were observed in the alignments. The original 'Low High Low High Low High' 3-step roll motif is clearly seen in the highest scoring A-Y alignment. Both sequence and A-Y score poorly in the leave-one-out cross validation, the most robust models being obtained with A-D. This suggests that A-Y may be overfitting the data with its large number of model parameters. Perfect recall results were obtained with all structural models and near perfect for sequence. This could be due to a poor test method with the random sequences always being very different from those of the dataset despite mono- or di-nucleotide shuffling. Real DNA datasets should be analysed before coming to any solid conclusions about what model settings are the absolute best. HMMs of four DNA protein binding site datasets are now produced and analysed in Chapter 9.

Chapter 9:
HMMs Of Four DNA Protein Binding Sites

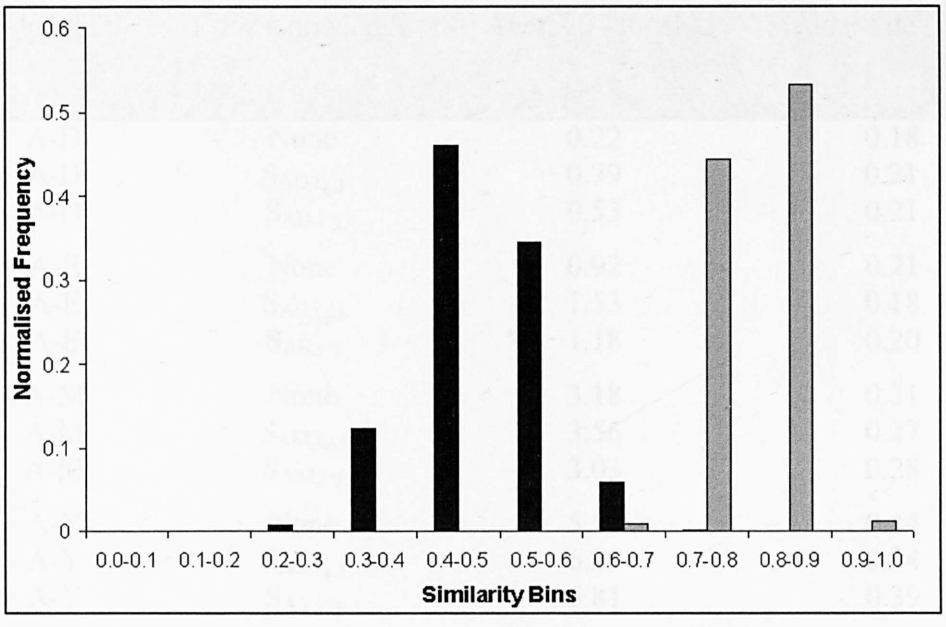
Four DNA protein binding site datasets are investigated by both sequence and structural (3-step roll) HMMs. Model performance is assessed by the three techniques introduced in Chapter 8: the non-validated approach, leave-one-out cross validation (LOO CV) and test set validation. Each dataset's size and diversity is analysed as a prerequisite to HMM generation, in order to increase our understanding of performance and check for biased redundant sequence information.

9.1. PrrA binding DNA

The PrrA binding site dataset was obtained from the Department of Molecular Biology and Biotechnology at Sheffield University (Laguri et al., 2003) in the hope that some further light could be shed upon the structural properties common to the DNA sequences. This collection of sequences bind to the effector domain of the PrrA protein of Rhodobacter sphaeroides, a proteo bacterium. PrrA plays an important role in the expression of genes involved in controlling metabolic changes between aerobic and anaerobic conditions. It has a helix-turn-helix motif that forms a dimer when activated. A DNA sequence alignment has been previously produced (Laguri et al., 2003) and identifies a consensus of inverted repeats separated by a variable spacing of three to nine nucleotides (GCGNC...GNCGC, where N means any nucleotide). Both the flexibility of the DNA and of a protein subunit are thought to be important factors in the underlying binding mechanism.

There are 38 sequences with an average length of 23.58 nucleotides, ranging from 20 to 26. The average pairwise Needleman Wunsch sequence similarity (Needleman and Wuncsh, 1970) is 0.48 and the analogous average structural similarity with respect to 3-step roll is 0.80. The PrrA binding sites are therefore more similar and less diverse by their structure than their sequence (Figure 9.1).

Figure 9.1: *Diversity of the PrrA binding site dataset with respect to Needleman-Wunsch sequence (black) and structure (grey) similarity distributions of all possible pairs.*



9.1.1. Non-validated analysis

A non-validated assessment of model performance was carried out as discussed in Chapter 8 section 8.6.1. The highest average model score was obtained with the A-Y alphabet using the $S_{AY,i_{\leftarrow j}}$ substitution matrix, closely followed by sequence (Table 9.1 and Figure 9.2). The model scores of A-Y with $S_{AY,i_{\leftarrow j}}$ are more precise than those of sequence, with a lower standard deviation. As alphabet size increases, the model scores tend to increase. $S_{i_{\leftarrow j}}$ is the highest scoring level of prior knowledge for all four structural alphabets. Surprisingly, in the A-M models no prior knowledge appears to be performing better than the $S_{i=j}$ matrix.

The highest scoring sequence model and highest scoring A-Y with $S_{AY,i_{\leftarrow j}}$ model were used to construct a sequence alignment (Figure 9.3a) and a structural alignment (Figure 9.3c) respectively. In the sequence alignment gaps (shown in black) are concentrated in the centre of the sequences and correspond to the variable spacing identified in a previous consensus pattern (Laguri et al., 2003). To either side of the gap area there are alternating columns of mainly orange (G) and blue (C), clearly reflected by the tallest letter stacks of the sequence logo (Figure 9.3c).

Table 9.1: *Non-validated scores for the PrrA binding site dataset with varying alphabet type and prior knowledge. Each average model score is taken over 100 models.*

Alphabet	Prior Knowledge	Average Model Score	Standard deviation
A-D	None	0.22	0.18
A-D	$S_{AD,i\leftarrow j}$	0.79	0.21
A-D	$S_{AD,i=j}$	0.53	0.21
A-E	None	0.92	0.21
A-E	$S_{AE,i\leftarrow j}$	1.53	0.18
A-E	$S_{AE,i=j}$	1.18	0.20
A-M	None	3.18	0.31
A-M	$S_{AM,i\leftarrow j}$	3.56	0.27
A-M	$S_{AM,i=j}$	3.03	0.28
A-Y	None	5.58	0.44
A-Y	$S_{AY,i\leftarrow j}$	6.29	0.34
A-Y	$S_{AY,i=j}$	5.81	0.39
Sequence	None	6.28	0.41

Figure 9.2: *The PrrA binding model score distributions for 100 non-validated HMM runs across the different alphabets with differing levels of prior knowledge. The different alphabets are sequence (orange), A-D (dark blue), A-E (red), A-M (green) and A-Y (light blue). The type of bar shading refers to the different levels of prior knowledge: no prior knowledge (solid shading), $S_{i\leftarrow j}$ substitution matrix (medium shading) and $S_{i=j}$ substitution matrix (light shading).*

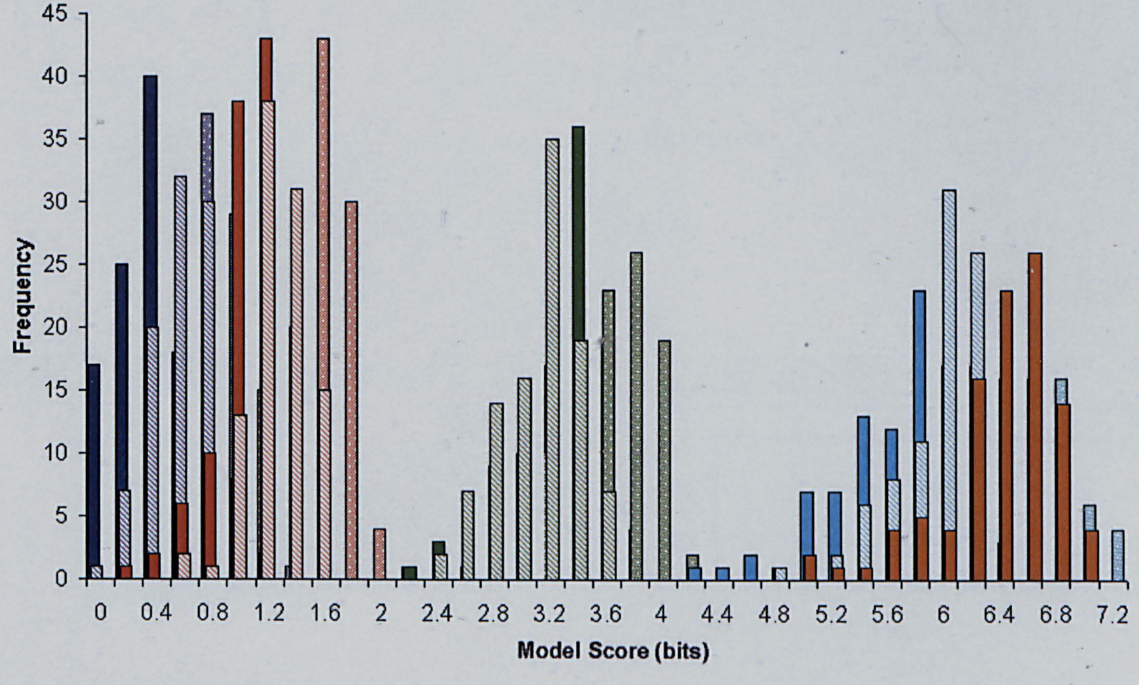
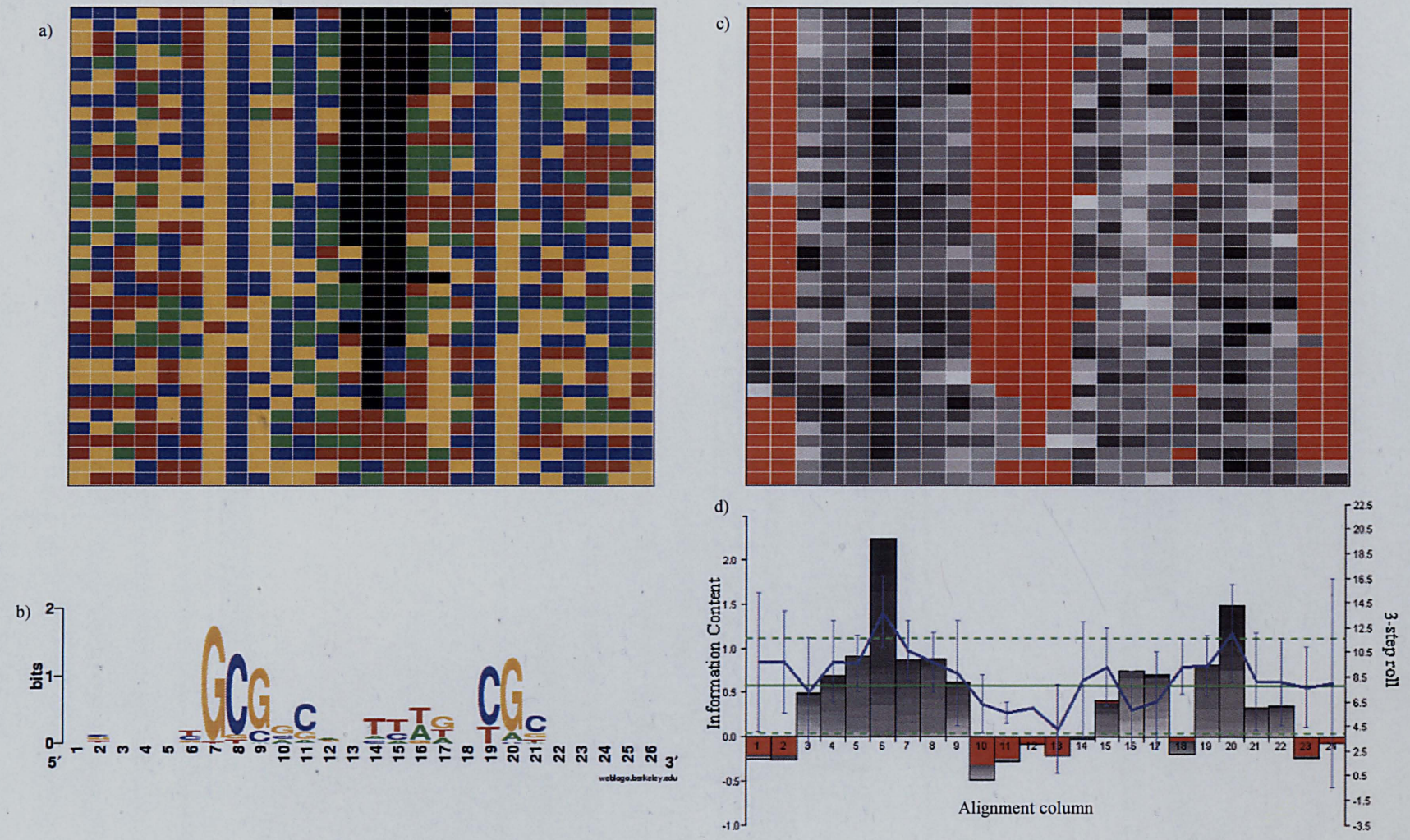


Figure 9.3: *Sequence alignment versus structural alignment of PrrA dataset. a)Sequence matrix plot, where black = gap, blue = C, orange = G, green = A and red = T. b) Sequence logo, obtained from Weblogo (Crooks et al., 2004). c) Structural matrix plot. Red = gap and light to dark shading is low to high 3-step roll. d) Structural logo*



The logo's consensus displays the importance of GCG at alignment positions 7, 8 and 9 and CGC at alignment positions 19, 20 and 21. The structural alignment and logo (Figure 9.3 c and d) were constructed and used to determine any structural implications with respect to the 3-step roll of these sequences. Again, a central region of variable spacing can be seen by gaps (shown in red) in the alignment (Figure 9.3 c). They suppress the heights of bars in the middle of the structural logo (Figure 9.3 d). Gaps also dominate the outer alignment columns 1, 2, 18, 23 and 24. The most important alignment position is 6 followed by 20, where in both cases the consensus 3-step roll is greater than average.

9.1.2. Leave-one-out cross validation

LOO CV was performed on the PrrA dataset (Chapter 8 section 8.6.2), where 100 models were made in the absence of each sequence. The predictive ability of all the structural models is poor with negative LOO CV scores (Table 9.2), meaning that on average a sequence removed from the training data will fit better to the null hypothesis (random DNA) than to the family of sequences to which it belongs. The percentage score reduction from the non-validated model scores to the LOO CV scores is greater than 100% (Table 9.2) for all the structural alphabets, corresponding to the negative LOO CV scores. The greatest percentage score reduction is for the A-D alphabet models that lack any prior knowledge. Although the non-validated A-Y models were marginally better than sequence, this LOO CV analysis shows that the sequence models have a much higher predictive ability than structure for the PrrA binding sites.

9.1.3. Test set validation

Test set validation was carried out as previously described (Chapter 8 section 8.6.3). Seven sequences (approximately 20% of the PrrA data) were randomly selected 100 times, in order to form the actives of 100 test sets. Nine inactive sequences were then generated for every active. Initially, model recall ability across all the alphabets with differing levels of prior knowledge and using the random method for generating inactives was considered (Table 9.3 and Figure 9.4). Sequence has the highest and

nearly perfect recall ability ($Recall_{NORM}$ of 0.986). However, structure is not far behind with values of $Recall_{NORM}$ varying from 0.830 (A-D with no prior knowledge) to 0.927 (A-Y with $S_{AY,i_{\leftarrow j}}$). Within each structural alphabet $S_{i_{\leftarrow j}}$ is again the best choice of prior knowledge.

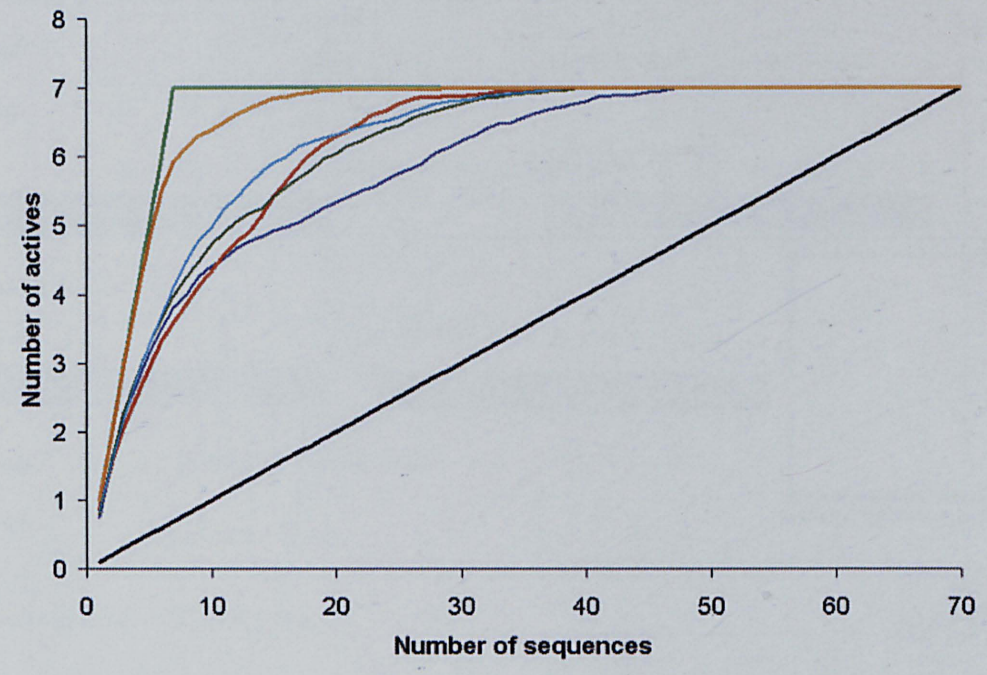
Table 9.2: *LOO CV scores for the PrrA binding site dataset with varying alphabet type and prior knowledge. Note that % score reduction refers to the reduction from the average non-validated model score to the analogous average LOO CV model score.*

Alphabet	Prior knowledge	LOO CV Score	% Score reduction
A-D	None	-1.68	864
A-D	$S_{AD,i_{\leftarrow j}}$	-0.85	208
A-D	$S_{AD,i=j}$	-0.915	273
A-E	None	-0.49	153
A-E	$S_{AE,i_{\leftarrow j}}$	-0.58	138
A-E	$S_{AE,i=j}$	-0.55	147
A-M	None	-0.87	127
A-M	$S_{AM,i_{\leftarrow j}}$	-0.84	124
A-M	$S_{AM,i=j}$	-0.67	122
A-Y	None	-1.92	134
A-Y	$S_{AY,i_{\leftarrow j}}$	-1.49	124
A-Y	$S_{AY,i=j}$	-1.45	125
Sequence	None	3.57	43

Table 9.3: *Values of Average $Recall_{NORM}$ for the PrrA binding site dataset with varying alphabet type and prior knowledge. N.B. (The inactives are generated using the random method).*

Alphabet	Prior knowledge	Average $Recall_{Norm}$
A-D	None	0.830
A-D	$S_{AD,i_{\leftarrow j}}$	0.873
A-D	$S_{AD,i=j}$	0.859
A-E	None	0.886
A-E	$S_{AE,i_{\leftarrow j}}$	0.914
A-E	$S_{AE,i=j}$	0.908
A-M	None	0.866
A-M	$S_{AM,i_{\leftarrow j}}$	0.913
A-M	$S_{AM,i=j}$	0.904
A-Y	None	0.862
A-Y	$S_{AY,i_{\leftarrow j}}$	0.927
A-Y	$S_{AY,i=j}$	0.913
Sequence	None	0.986

Figure 9.4: *Cumulative recall plot for the PrrA binding site dataset with varying alphabet type. Sequence is orange, A-D is dark blue, A-E is red, A-M is dark green, A-Y is light blue, random recall is black and ideal recall is bright green. N.B. (The inactives are generated using the random method).*



The effect that the method for generating the inactives has upon recall was explored for the A-Y alphabet with $S_{AY,i\leftarrow j}$ and the traditional sequence alphabet (Table 9.4 and Figure 9.5). The expected degrading effect upon recall when going from random to mono shuffled to doublet shuffled inactives was observed. In all cases the sequence recall remains higher than structure. The active and inactive score distributions were plotted (Figure 9.6) and T-tests assessing their separation were performed (Table 9.5).

Table 9.4: *Exploring the effect different methods for generating the inactives have upon values of* $Recall_{NORM}$ *for the sequence and A-Y (with* $S_{AY,i\leftarrow j}$*) models for the PrrA binding site dataset.*

Inactive Class	Sequence $Recall_{NORM}$	A-Y $Recall_{NORM}$
Random	0.986	0.927
Mono shuffling	0.963	0.885
Dinucleotide shuffling	0.938	0.849

Figure 9.5: *Cumulative recall plot for the PrrA binding site dataset, considering the sequence models (black) and the A-Y roll alphabet models (blue). Ideal recall is shown in green and random recall in red. The $S_{i \swarrow j}$ substitution matrix is used for structure and the three classes of inactives were explored: the random inactive (solid lines), mono shuffled inactives (dashed lines) and doublet shuffled (lighter shaded lines).*

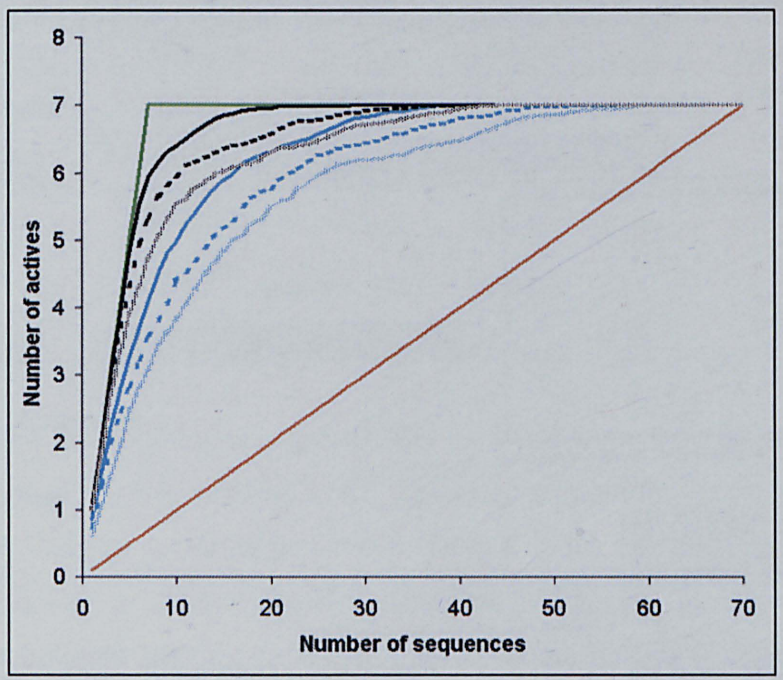


Figure 9.6: *PrrA score distributions of the sequence inactives (red), sequence actives (green), structure inactives (purple) and structure actives (light blue). The level of shading refers to the method used to generate the inactives: random generation (solid lines), mono shuffling (dashed lines) and doublet shuffling (lighter shaded lines). The $S_{i \swarrow j}$ substitution matrix is used for structure.*

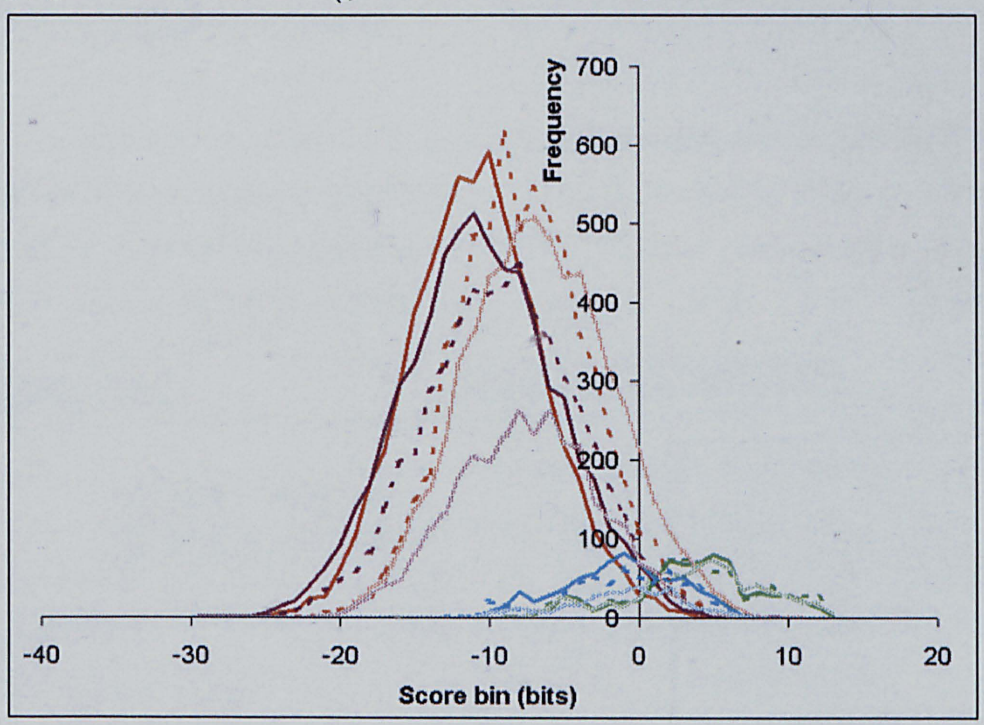


Table 9.5: *T-test analyses between active and inactive score distributions from Figure 9.6. DOF is the number of degrees of freedom.* $\bar{x}_1 - \bar{x}_2$ *is the mean separation between the two distributions. The* $S_{i \leftarrow j}$ *substitution matrix is used for structure.*

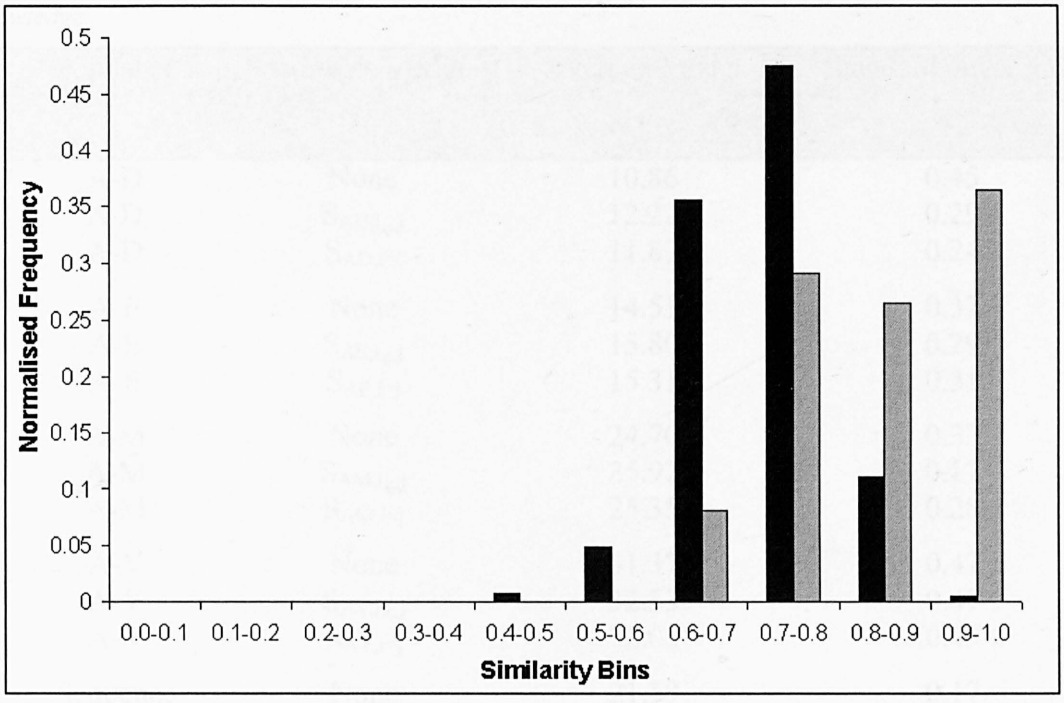
Alphabet	Type of inactives	DOF	T	$\bar{x}_1 - \bar{x}_2$
Sequence	Random	867	82.06	14.35
Sequence	Mono	850	63.60	11.55
Sequence	Doublet	889	56.99	10.44
A-Y	Random	998	58.95	9.32
A-Y	Mono	839	43.82	7.84
A-Y	Doublet	459	27.23	6.32

All active score distributions are significantly different from their inactive analogues (values of T being much larger than 2.58). The mean separation is greater for sequence than structure (leading to the higher recall ability). The separation decreases when going from random to doublet shuffled inactives. Although the majority of active sequences are being scored negatively with the structural models (Figure 9.6) they are still being scored significantly less negatively than their inactive counterparts. This explains why the LOO CV scores are so low, but the recall ability is high.

9.2. PPARg Factor Binding Sites

This dataset was obtained from a list of 95 transcription factors, each with at least 20 associated binding sites (Barash et al., 2003), the origin of the data being the TRANSFAC database (Wingender et al., 2001). The factor concerned is a peroxisome proliferator-activated receptor of type gamma (PPARg) that binds to DNA via two zinc fingers. This nuclear receptor induces proliferation of peroxisomes in the cytoplasm, a peroxisome's function being to oxidise materials in the cell and then catalyse the destruction of the resulting poisonous hydrogen peroxide by-product. An excellent review of PPARs is present in the literature (Berger and Moller, 2002). There are 72 sequences with an average length of 19.44, ranging from 15 to 21. The average sequence similarity is 0.71 and structural similarity 0.84 (Figure 9.7).

Figure 9.7: *Diversity of the PPARg binding site dataset with respect to Needleman Wunsch sequence (black) and structure (grey) similarity distributions of all possible pairs.*



9.2.1. Non-validated analysis

The models built from the A-Y alphabet with $S_{AY,i\leftarrow j}$ are the highest scoring (Table 9.6). All non-validated A-M and A-Y model scores beat those of sequence, regardless of the level of prior knowledge used (Figure 9.8). However, sequence does have the smallest standard deviation, therefore it has the most reproducible alignment solutions. The model scores increase with the size of the structural alphabet and are the highest for the $S_{i\leftarrow j}$ matrices followed by $S_{i=j}$ matrices.

A very strong sequence consensus (GGTCAAAGGTCA) and structural consensus (repeating low to high 3-step roll) have been identified from the top scoring sequence and structure model. Clear vertical bands of colour are present in the matrix plots (Figure 9.9a and c). The letter stacks in the sequence logo are high and dominated by single letters (Figure 9.9b). The structural logo has high information content with large bar heights and negligible 3-step roll variance (Figure 9.9d). Gaps are present in both alignments, but always towards the edges. This is particularly the case for structure, where gaps are purely at the ends and never interrupt a sequence.

Table 9.6: *Non-validated scores for the PPARg binding site dataset with varying alphabet type and prior knowledge.*

Alphabet	Prior Knowledge	Average Model Score	Standard deviation
A-D	None	10.86	0.45
A-D	$S_{AD,i \leftarrow j}$	12.22	0.29
A-D	$S_{AD,i=j}$	11.87	0.24
A-E	None	14.51	0.32
A-E	$S_{AE,i \leftarrow j}$	15.80	0.29
A-E	$S_{AE,i=j}$	15.31	0.31
A-M	None	24.70	0.32
A-M	$S_{AM,i \leftarrow j}$	25.92	0.17
A-M	$S_{AM,i=j}$	25.35	0.28
A-Y	None	31.37	0.47
A-Y	$S_{AY,i \leftarrow j}$	32.33	0.47
A-Y	$S_{AY,i=j}$	32.05	0.45
Sequence	None	21.59	0.17

Figure 9.8: *PPARg binding model score distributions for 100 non-validated HMM runs across the different alphabets with differing levels of prior knowledge. The different alphabets are sequence (orange), A-D (dark blue), A-E (red), A-M (green) and A-Y (light blue). The type of bar shading refers to the different levels of prior knowledge: no prior knowledge (solid shading), $S_{i \leftarrow j}$ substitution matrix (medium shading) and $S_{i=j}$ substitution matrix (light shading).*

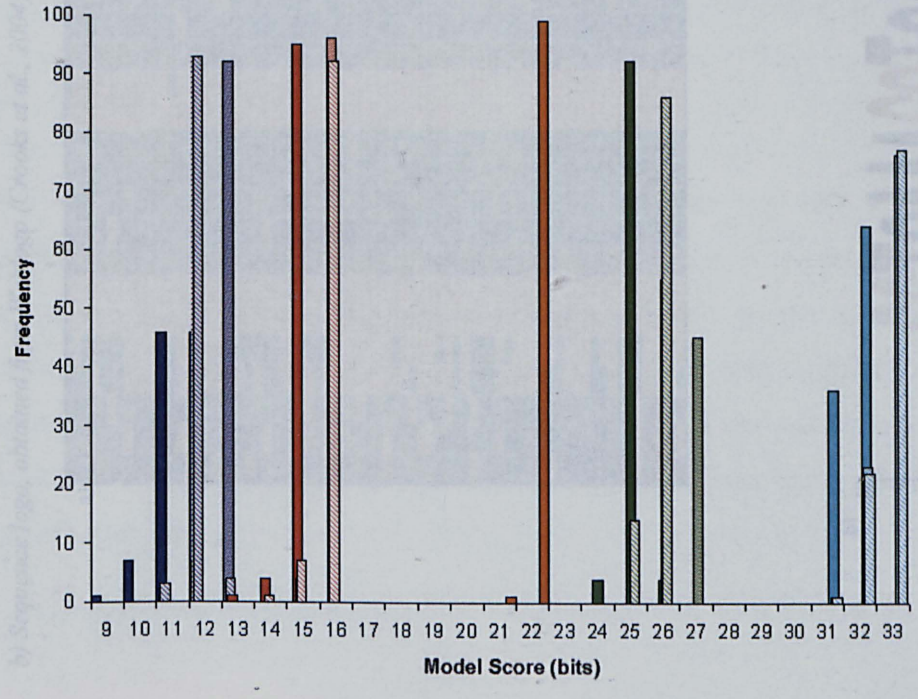
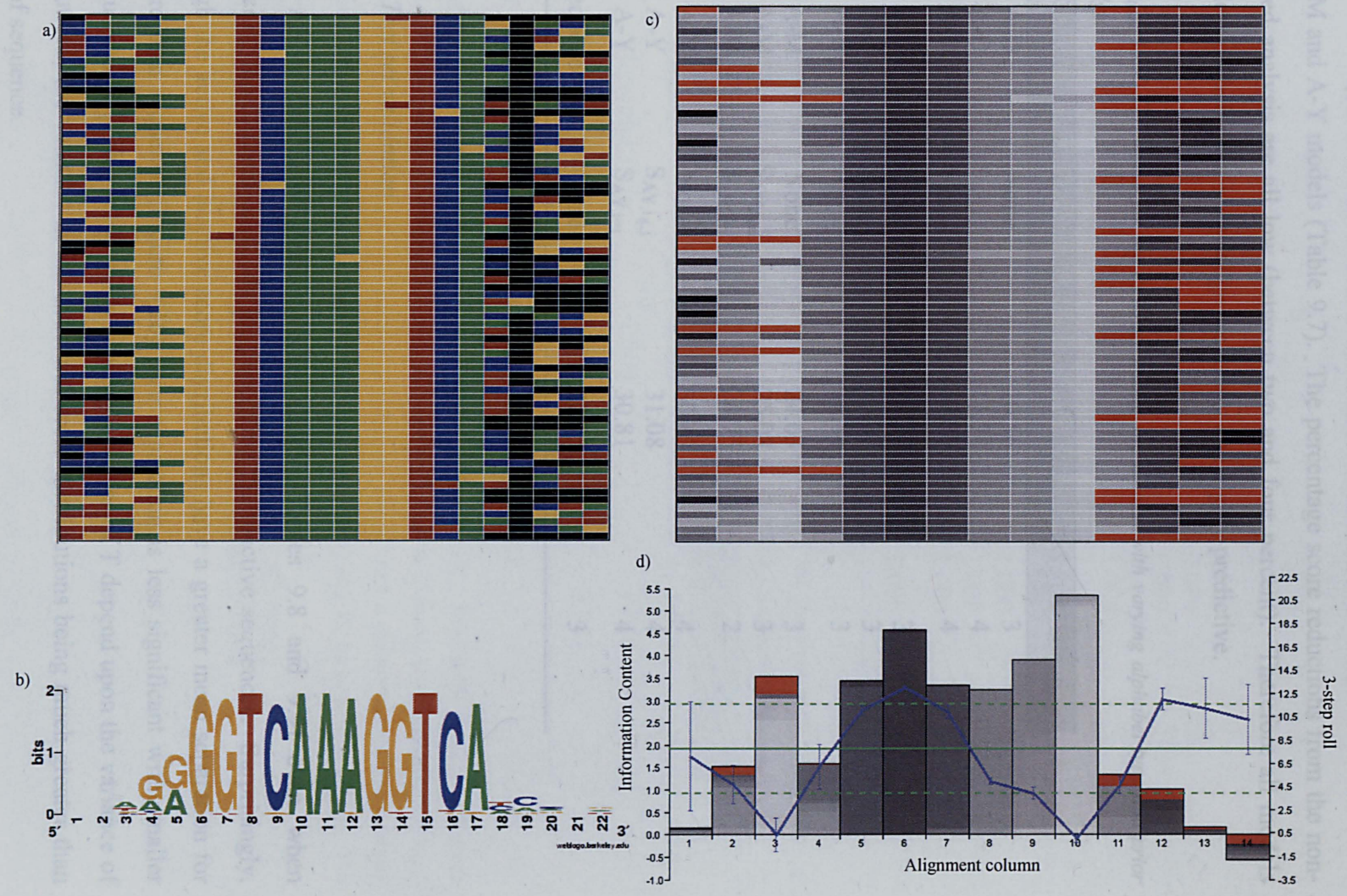


Figure 9.9: *Sequence alignment versus structural alignment of PPARg dataset. a)Sequence matrix plot, where black = gap, blue = C, orange = G, green = A and red = T. b) Sequence logo, obtained from Weblogo (Crooks et al., 2004). c) Structural matrix plot. Red = gap and light to dark shading is low to high 3-step roll. d) Structural logo*



9.2.2. *Leave-one-out cross validation*

A-Y $S_{i \leftarrow j}$ is the highest LOO CV model scorer with sequence still falling behind the A-M and A-Y models (Table 9.7). The percentage score reductions from the non-validated analysis are all low (between two and four percent). Therefore all models regardless of their prior knowledge are robust and highly predictive.

Table 9.7: LOO CV *scores for the PPARg binding site dataset with varying alphabet type and prior knowledge.*

Alphabet	Prior knowledge	LOO CV Score	% Score reduction
A-D	None	10.56	3
A-D	$S_{AD,i \leftarrow j}$	11.74	4
A-D	$S_{AD,i=j}$	11.39	4
A-E	None	14.10	3
A-E	$S_{AE,i \leftarrow j}$	15.37	3
A-E	$S_{AE,i=j}$	14.84	3
A-M	None	24.03	3
A-M	$S_{AM,i \leftarrow j}$	25.08	3
A-M	$S_{AM,i=j}$	24.75	2
A-Y	None	30.10	4
A-Y	$S_{AY,i \leftarrow j}$	31.08	4
A-Y	$S_{AY,i=j}$	30.81	4
Sequence	None	20.95	3

9.2.3. *Test set validation*

All models have perfect recall ability (Tables 9.8 and 9.9) even when considering the different methods for generating the inactive sequences. Surprisingly, although the active and inactive score distributions have a greater mean separation for structure than sequence (Figure 9.10) their separation is less significant with smaller magnitudes of T (Table 9.10). This is because values of T depend upon the variance of distributions, the variance in the structural score distributions being much greater than those of sequence.

Table 9.8: *Values of Average $Recall_{NORM}$ for the PPARg binding site dataset with varying alphabet type and prior knowledge. N.B. (The inactives are generated using the random method).*

Alphabet	Prior knowledge	Average $Recall_{Norm}$
A-D	None	0.997
A-D	$S_{AD,i\leftarrow j}$	0.999
A-D	$S_{AD,i=j}$	0.999
A-E	None	0.996
A-E	$S_{AE,i\leftarrow j}$	0.999
A-E	$S_{AE,i=j}$	0.999
A-M	None	1.000
A-M	$S_{AM,i\leftarrow j}$	1.000
A-M	$S_{AM,i=j}$	1.000
A-Y	None	1.000
A-Y	$S_{AY,i\leftarrow j}$	1.000
A-Y	$S_{AY,i=j}$	1.000
Sequence	None	1.000

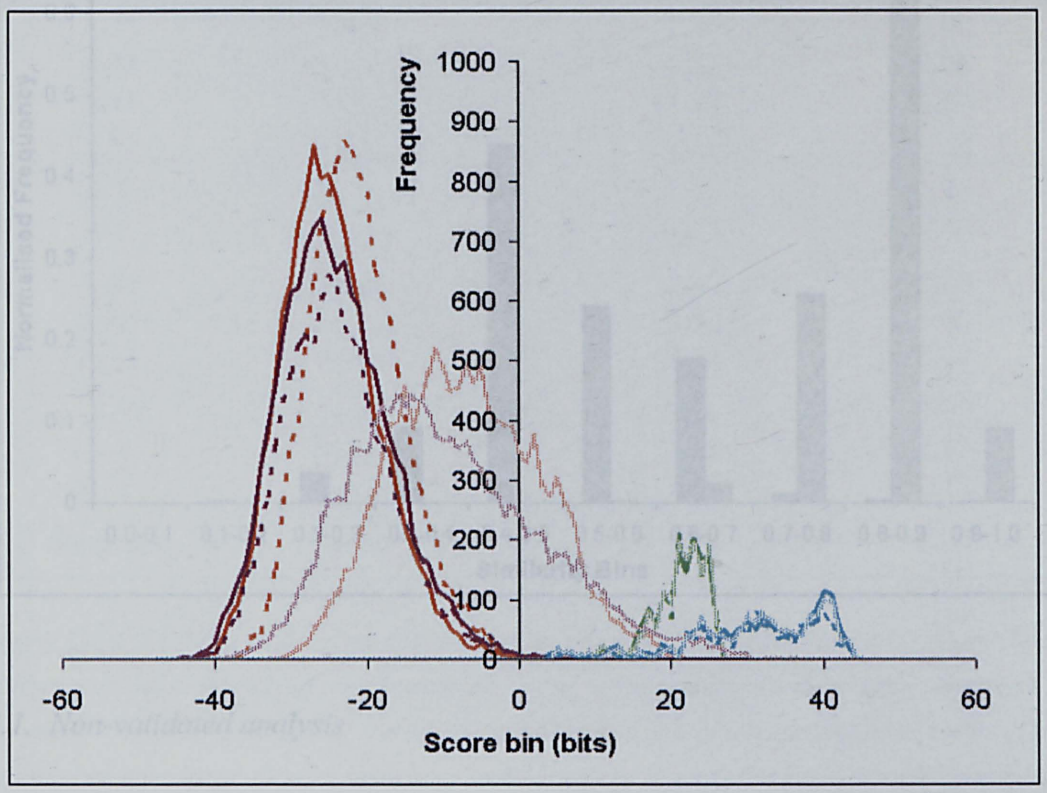
Table 9.9: *Exploring the effect different methods for generating the inactives have upon values of $Recall_{NORM}$ for the sequence and A-Y (with $S_{AY,i\leftarrow j}$) models for the PPARg binding site dataset.*

Inactive Class	Sequence $Recall_{NORM}$	A-Y $Recall_{NORM}$
Random	1.000	1.000
Mono shuffling	1.000	1.000
Dinucleotide shuffling	0.997	0.991

Table 9.10: T-test analysis upon the score distributions of Figure 9.10. DOF is the number of degrees of freedom. The samples are labelled in 3 parts separated by '_'. The 1st part is the alphabet type: 'sequ' means sequence and 'AY' means A-Y 3-step roll alphabet. The 2nd part is the method used to generate the inactives: 'r' means random generation, 'm' is for mono shuffling and 'd' for doublet shuffling. The 3rd part is 'a' for an active distribution and 'i' for an inactive distribution.

Alphabet	Type of inactives	DOF	T	$\bar{x}_1 - \bar{x}_2$
Sequence	Random	2646	428.92	46.03
Sequence	Mono	2524	397.06	42.48
Sequence	Doublet	4571	225.86	28.23
A-Y	Random	1607	223.50	55.39
A-Y	Mono	1310	196.05	54.06
A-Y	Doublet	2048	153.06	40.52

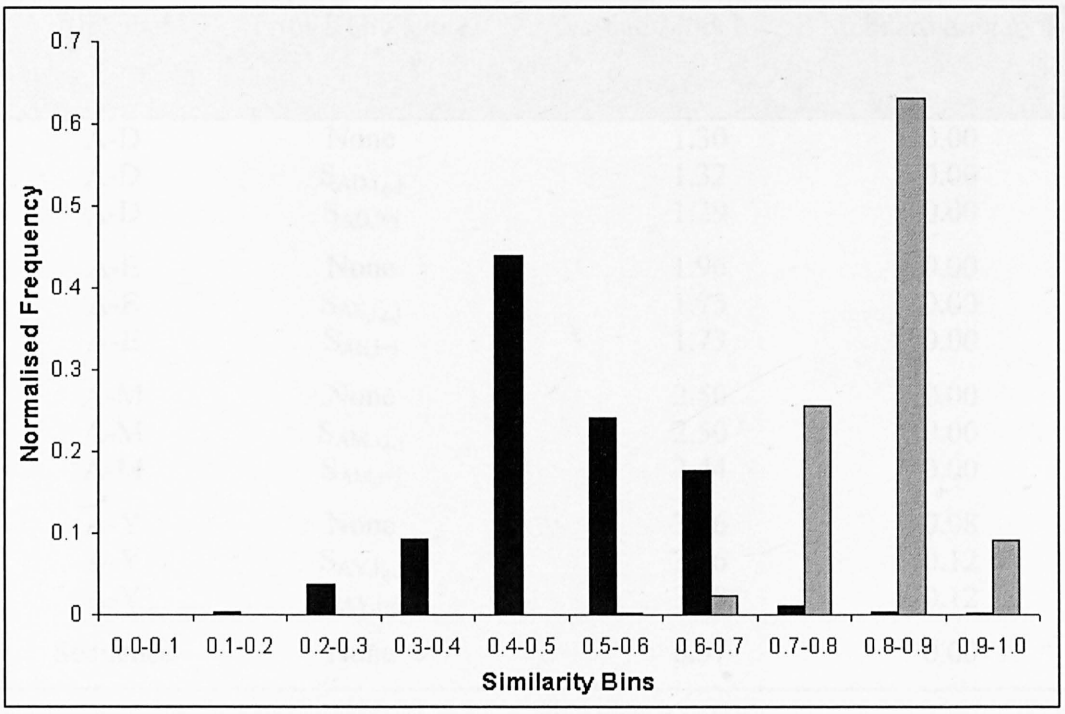
Figure 9.10: *PPARg score distributions of the sequence inactives (red), sequence actives (green), structure inactives (purple) and structure actives (light blue). The level of shading refers to the method used to generate the inactives: random generation (solid lines), mono shuffling (dashed lines) and doublet shuffling (lighter shading). The $S_{i \leftarrow j}$ substitution matrix is used for structure.*



9.3. FIS Binding Sites

This dataset is a collection of FIS binding sites pre-aligned by their sequence (Ussery et al., 2001). The protein concerned is a Factor for Inversion Stimulation (FIS) associated with E.Coli chromatin. It possesses a weak consensus DNA binding site and has a helix-turn-helix motif. FIS is a gene expression regulator that stabilises DNA looping (Ussery et al., 2001). Ten redundant sequences were found in the data and removed. 91 aligned sequences of length 15 nucleotides remained. The average sequence similarity is 0.48 and the average structural similarity is 0.83 (Figure 9.11).

Figure 9.11: *Diversity of the FIS binding site dataset with respect to Needleman Wunsch sequence (black) and structure (grey) similarity distributions of all possible pairs.*



9.3.1. Non-validated analysis

The sequence models have the highest scores (Table 9.11 and Figure 9.12). The average model scores increase with the structural alphabet size, but the level of prior knowledge used within each alphabet does not have a considerable effect on the model performance. The A-D, A-E, A-M and sequence alphabet models all have standard deviation of zero, corresponding to one unique alignment solution. The A-Y alphabet also has very precise model scores with a standard deviation of only 0.1 bits.

The majority of FIS sites start with guanine and end with cytosine and have a high density of adenine and thymine in the centre (Figure 9.13a and b). No gaps are present in the sequence alignment, since the sequences are pre-aligned and of identical length. No striking structural patterns can be seen in the 3-step roll alignment and logo plot (Figure 9.13c and d). This suggests that roll is not an important degree of freedom in FIS binding site recognition. Note the gaps present in alignment column 4. These must refer to an insert state being used by a single sequence.

Table 9.11: *Non-validated scores for the FIS binding site dataset with varying alphabet type and prior knowledge.*

Alphabet	Prior Knowledge	Average Model Score	Standard deviation
A-D	None	1.30	0.00
A-D	$S_{AD,i \leftarrow j}$	1.32	0.00
A-D	$S_{AD,i=j}$	1.29	0.00
A-E	None	1.96	0.00
A-E	$S_{AE,i \leftarrow j}$	1.75	0.00
A-E	$S_{AE,i=j}$	1.73	0.00
A-M	None	2.50	0.00
A-M	$S_{AM,i \leftarrow j}$	2.50	0.00
A-M	$S_{AM,i=j}$	2.44	0.00
A-Y	None	3.56	0.08
A-Y	$S_{AY,i \leftarrow j}$	3.56	0.12
A-Y	$S_{AY,i=j}$	3.48	0.12
Sequence	None	6.57	0.00

Figure 9.12: *FIS binding model score distributions for 100 non-validated HMM runs across the different alphabets with differing levels of prior knowledge. The different alphabets are sequence (orange), A-D (dark blue), A-E (red), A-M (green) and A-Y (light blue). The type of bar shading refers to the different levels of prior knowledge: no prior knowledge (solid shading), $S_{i \leftarrow j}$ substitution matrix (medium shading) and $S_{i=j}$ substitution matrix (light shading).*

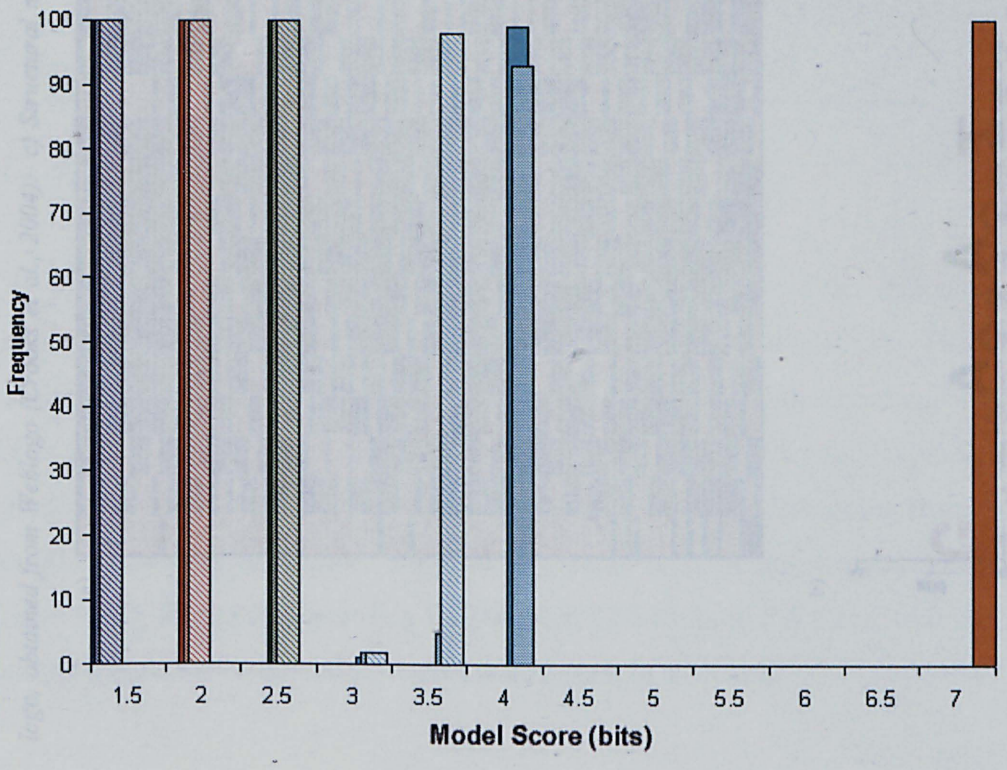
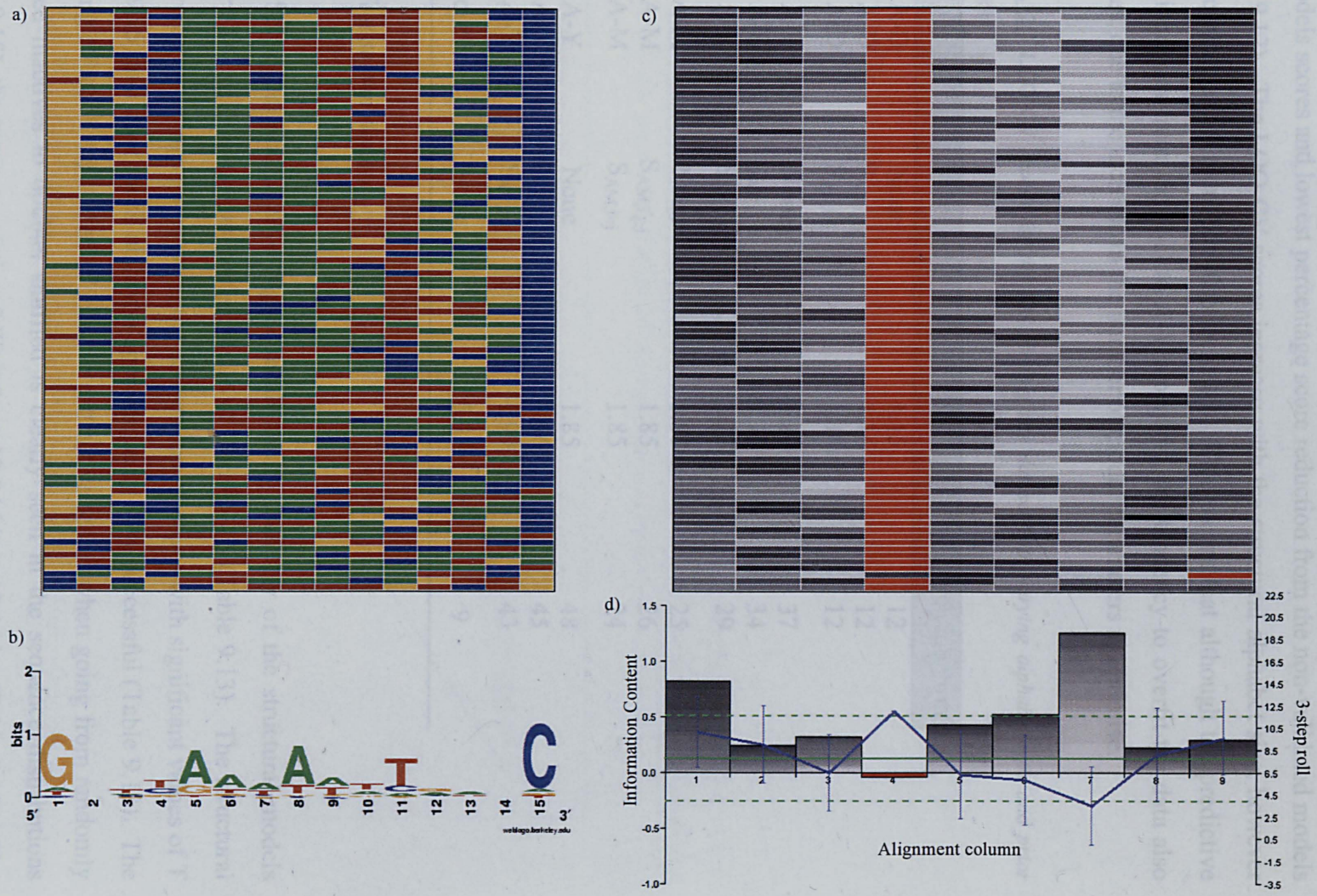


Figure 9.13: *Sequence alignment versus structural alignment of FIS dataset. a)Sequence matrix plot, where blue = C, orange = G, green = A and red = T. b) Sequence logo, obtained from Weblogo (Crooks et al., 2004). c) Structural matrix plot, where red = gap and light to dark shading is low to high 3-step roll. d) Structural logo*



9.3.2. *Leave-one-out validation*

The sequence models are the most predictive and robust with the highest LOO CV models scores and lowest percentage score reduction from the non-validated models (Table 9.12). The LOO CV scores increase with the structural alphabet size, however the percentage score reduction also increases. This means that although the predictive ability increases from the A-D to A-Y alphabets, the tendency to overfit the data also increases, due to the increase in the number of model parameters to optimise.

Table 9.12: LOO CV *scores for the FIS binding site dataset with varying alphabet type and prior knowledge.*

Alphabet	Prior knowledge	LOO CV Score	% Score reduction
A-D	None	1.15	12
A-D	$S_{AD,i \leftarrow j}$	1.16	12
A-D	$S_{AD,i=j}$	1.14	12
A-E	None	1.24	37
A-E	$S_{AE,i \leftarrow j}$	1.15	34
A-E	$S_{AE,i=j}$	1.22	29
A-M	None	1.87	25
A-M	$S_{AM,i \leftarrow j}$	1.85	26
A-M	$S_{AM,i=j}$	1.85	24
A-Y	None	1.85	48
A-Y	$S_{AY,i \leftarrow j}$	1.96	45
A-Y	$S_{AY,i=j}$	1.99	43
Sequence	None	5.97	9

9.3.3. *Test set validation*

Sequence models have higher recall ability than any of the structural models (Figure 9.14) with an almost perfect value of $Recall_{NORM}$ (Table 9.13). The structural models clearly distinguish between the actives and inactives with significant values of T (Table 9.15) and their recall ability is approximately 80% successful (Table 9.13). The degrading effect on separation of inactive and active scores when going from randomly generated inactives to doublet shuffled is clearly seen in the sequence distributions (Figure 9.15), the mean separation falling from 12.84 bits to 5.4 bits (Table 9.15). The analogous but smaller degrading effect is also seen in structure.

Table 9.13: *Values of Average Recall$_{NORM}$ for the FIS binding site dataset with varying alphabet type and prior knowledge. N.B. (The inactives are generated using the random method).*

Alphabet	Prior knowledge	Average Recall$_{Norm}$
A-D	None	0.806
A-D	$S_{AD,i\leftarrow j}$	0.805
A-D	$S_{AD,i=j}$	0.801
A-E	None	0.813
A-E	$S_{AE,i\leftarrow j}$	0.812
A-E	$S_{AE,i=j}$	0.810
A-M	None	0.842
A-M	$S_{AM,i\leftarrow j}$	0.840
A-M	$S_{AM,i=j}$	0.836
A-Y	None	0.831
A-Y	$S_{AY,i\leftarrow j}$	0.835
A-Y	$S_{AY,i=j}$	0.830
Sequence	None	0.971

Figure 9.14: *Cumulative recall plot for the FIS binding site dataset with varying alphabet type. A-D is dark blue, A-E is red, A-M is green, A-Y is light blue, sequence is orange, random recall is black and ideal recall is bright green. N.B. (The inactives are generated using the random method)*

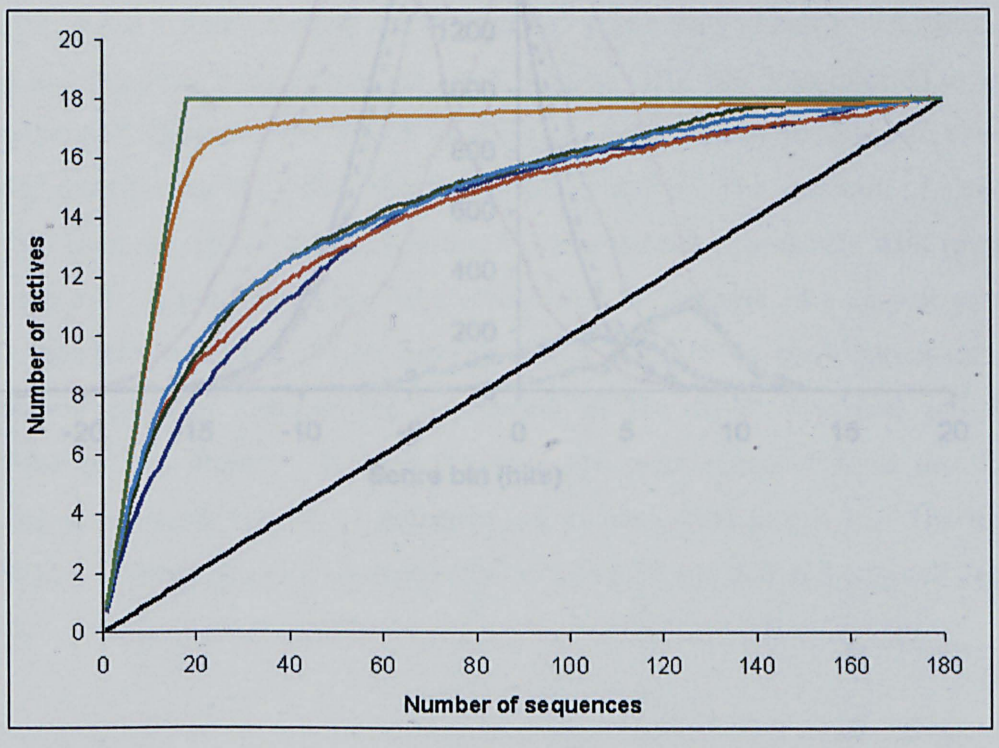


Table 9.14: *Exploring the effect different methods for generating the inactives have upon values of* $Recall_{NORM}$ *for the sequence and A-Y (with* $S_{AY,i \leftarrow j}$*)models for the FIS binding site dataset.*

Inactive Class	Sequence $Recall_{NORM}$	A-Y $Recall_{NORM}$
Random	0.971	0.835
Mono shuffling	0.965	0.816
Dinucleotide shuffling	0.850	0.781

Figure 9.15: *FIS score distributions of the sequence inactives (red), sequence actives (green), structure inactives (purple) and structure actives (light blue). The level of shading refers to the method used to generate the inactives: random generation (solid lines), mono shuffling (dashed lines) and doublet shuffling (lighter shading). The* $S_{i \leftarrow j}$ *substitution matrix is used for structure*

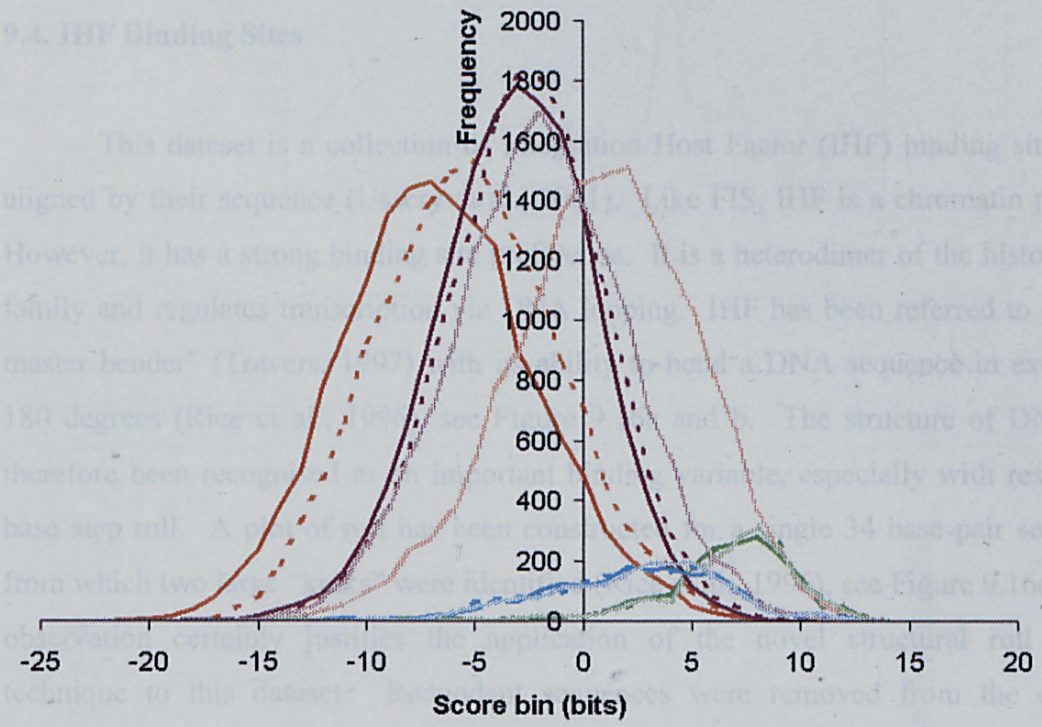


Table 9.15: *T-test analyses between active and inactive score distributions from Figure 9.15. DOF is the number of degrees of freedom.* $\bar{x}_1 - \bar{x}_2$ *is the mean separation between the two distributions. The* $S_{i \leftarrow j}$ *substitution matrix is used for structure.*

Alphabet	Type of inactives	DOF	T	$\bar{x}_1 - \bar{x}_2$
Sequence	Random	2303	126.8507	12.84
Sequence	Mono	2291	118.6777	11.7
Sequence	Doublet	2260	52.4688	5.4
A-Y	Random	2107	52.05861	5.33
A-Y	Mono	2104	47.93003	4.97
A-Y	Doublet	2194	42.80389	4.37

9.4. IHF Binding Sites

This dataset is a collection of Integration Host Factor (IHF) binding sites pre-aligned by their sequence (Ussery et al., 2001). Like FIS, IHF is a chromatin protein. However, it has a strong binding site preference. It is a heterodimer of the histone-like family and regulates transcription via DNA looping. IHF has been referred to as "the master bender" (Travers, 1997) with its ability to bend a DNA sequence in excess of 180 degrees (Rice et al., 1996), see Figure 9.16a and b. The structure of DNA has therefore been recognised as an important binding variable, especially with respect to base step roll. A plot of roll has been constructed for a single 34 base pair sequence from which two large "kinks" were identified (Rice et al., 1996), see Figure 9.16c. This observation certainly justifies the application of the novel structural roll HMM technique to this dataset. Redundant sequences were removed from the dataset, resulting in a dataset size of 47 sequences all of nucleotide length 15. The average Needleman-Wunsch pairwise sequence similarity is 0.53 and that of 3-step roll structure is 0.85.

Figure 9.16: *Complex of IHF bound to DNA (Rice et al., 1996). a) Front view. b) Top view. The two protein subunits are shown in white and pink with an identified DNA consensus highlighted in green. Intercalating prolines are shown in yellow. c) Roll profile (Rice et al.,1996). No points plotted for TTG, which are in non-Watson-Crick configurations.*

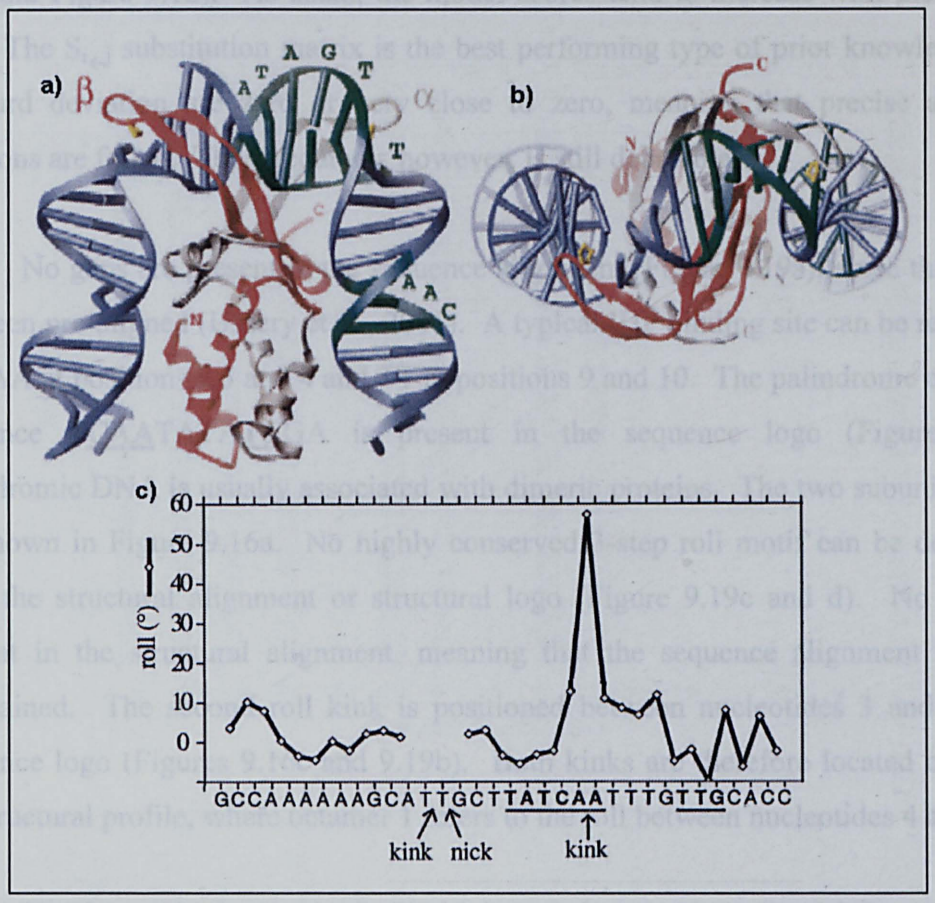
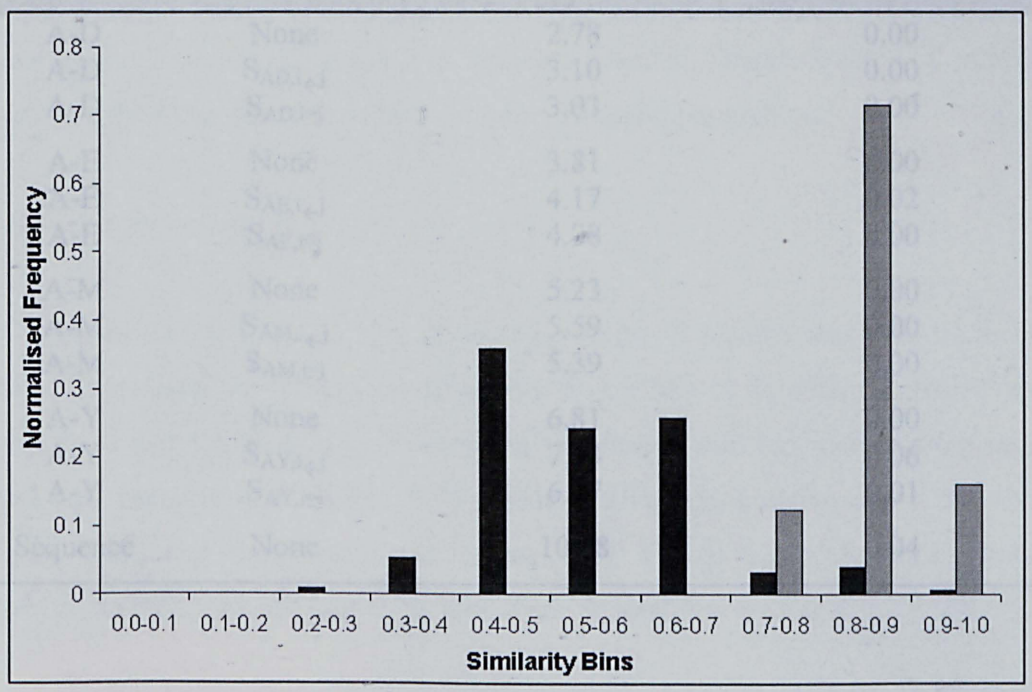


Figure 9.17: *Diversity of the IHF binding site dataset with respect to Needleman Wunsch sequence (black) and structure (grey) similarity distributions of all possible pairs.*



9.4.1. Non-validated analysis

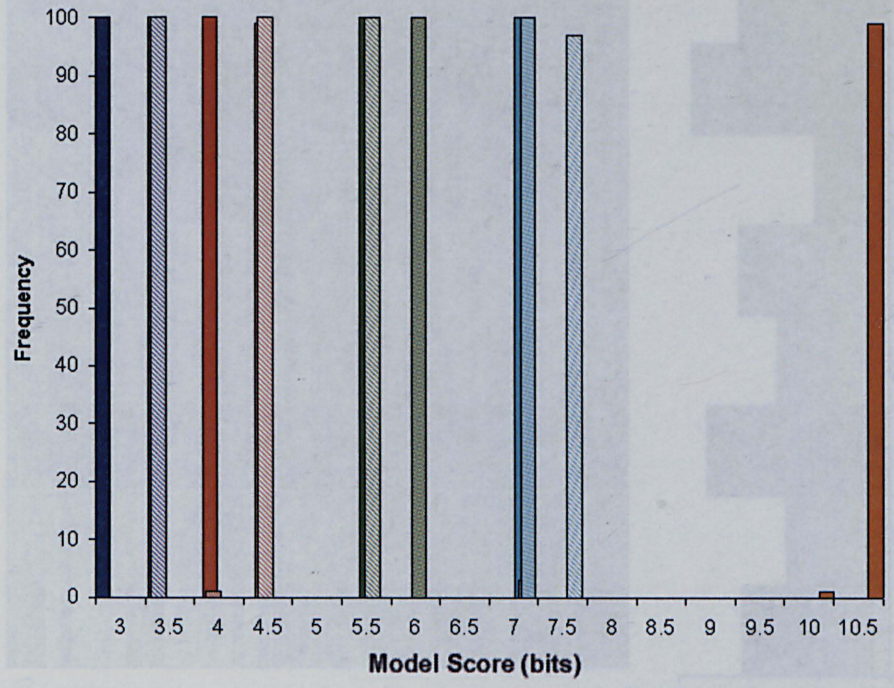
Sequence gives the highest scoring models in comparison to structure (Table 9.16 and Figure 9.18). As usual, the model scores tend to increase with the alphabet size. The $S_{i\leftarrow j}$ substitution matrix is the best performing type of prior knowledge. All standard deviation are zero or very close to zero, meaning that precise alignment solutions are found. Their accuracy, however, is still debatable.

No gaps are present in the sequence alignment (Figure 9.19a), since this dataset has been pre-aligned (Ussery et al., 2001). A typical IHF binding site can be recognised by CAA at positions 2,3 and 4 and TT at positions 9 and 10. The palindrome consensus sequence T**CAA**TATA**TT**GA is present in the sequence logo (Figure 9.19b). Palindromic DNA is usually associated with dimeric proteins. The two subunits of IHF are shown in Figure 9.16a. No highly conserved 3-step roll motif can be deciphered from the structural alignment or structural logo (Figure 9.19c and d). No gaps are present in the structural alignment, meaning that the sequence alignment has been maintained. The second roll kink is positioned between nucleotides 3 and 4 in the sequence logo (Figures 9.16c and 9.19b). Both kinks are therefore located outside of the structural profile, where octamer 1 refers to the roll between nucleotides 4 and 5.

Table 9.16: *Non-validated scores for the IHF binding site dataset with varying alphabet type and prior knowledge.*

Alphabet	Prior Knowledge	Average Model Score	Standard deviation
A-D	None	2.78	0.00
A-D	$S_{AD,i\leftarrow j}$	3.10	0.00
A-D	$S_{AD,i=j}$	3.03	0.00
A-E	None	3.81	0.00
A-E	$S_{AE,i\leftarrow j}$	4.17	0.02
A-E	$S_{AE,i=j}$	4.08	0.00
A-M	None	5.23	0.00
A-M	$S_{AM,i\leftarrow j}$	5.59	0.00
A-M	$S_{AM,i=j}$	5.39	0.00
A-Y	None	6.81	0.00
A-Y	$S_{AY,i\leftarrow j}$	7.14	0.06
A-Y	$S_{AY,i=j}$	6.97	0.01
Sequence	None	10.28	0.04

Figure 9.18: *IHF binding model score distributions for 100 non-validated HMM runs across the different alphabets with differing levels of prior knowledge. The different alphabets are sequence (orange), A-D (dark blue), A-E (red), A-M (green) and A-Y (light blue). The type of bar shading refers to the different levels of prior knowledge: no prior knowledge (solid shading), $S_{i\neq j}$ substitution matrix (medium shading) and $S_{i=j}$ substitution matrix (light shading)*



9.4.2. Leave-one-out validation

The sequence models are the most predictive with the highest LOO CV model scores and the most robust with the smallest percentage score reduction from the non-validated model scores (Table 9.17). The A-M 3-step roll alphabet models have a higher predictive ability than the A-Y alphabet models. Therefore the A-Y models are overfitting the data, due to their large number of model parameters.

9.4.3. Test set validation

Surprisingly, even though no strong 3-step roll pattern was visualised, all the structural models have excellent recall ability with values of $Recall_{NORM}$ greater than 0.9 (Table 9.18 and Figure 9.20). Even when the double shuffled inactives are used the recall still remains high (Table 9.19). The active and inactive score distributions (Figure 9.21) and their associated values of T (Table 9.20) show discriminative capability in structure and sequence, with mean separations greater than 10 bits.

Figure 9.19: *Sequence alignment versus structural alignment of IHF dataset. a)Sequence matrix plot, where blue = C, orange = G, green = A and red = T. b) Sequence logo, obtained from Weblogo (Crooks et al., 2004). c) Structural matrix plot, where red = gap and light to dark shading is low to high 3-step roll. d) Structural logo*

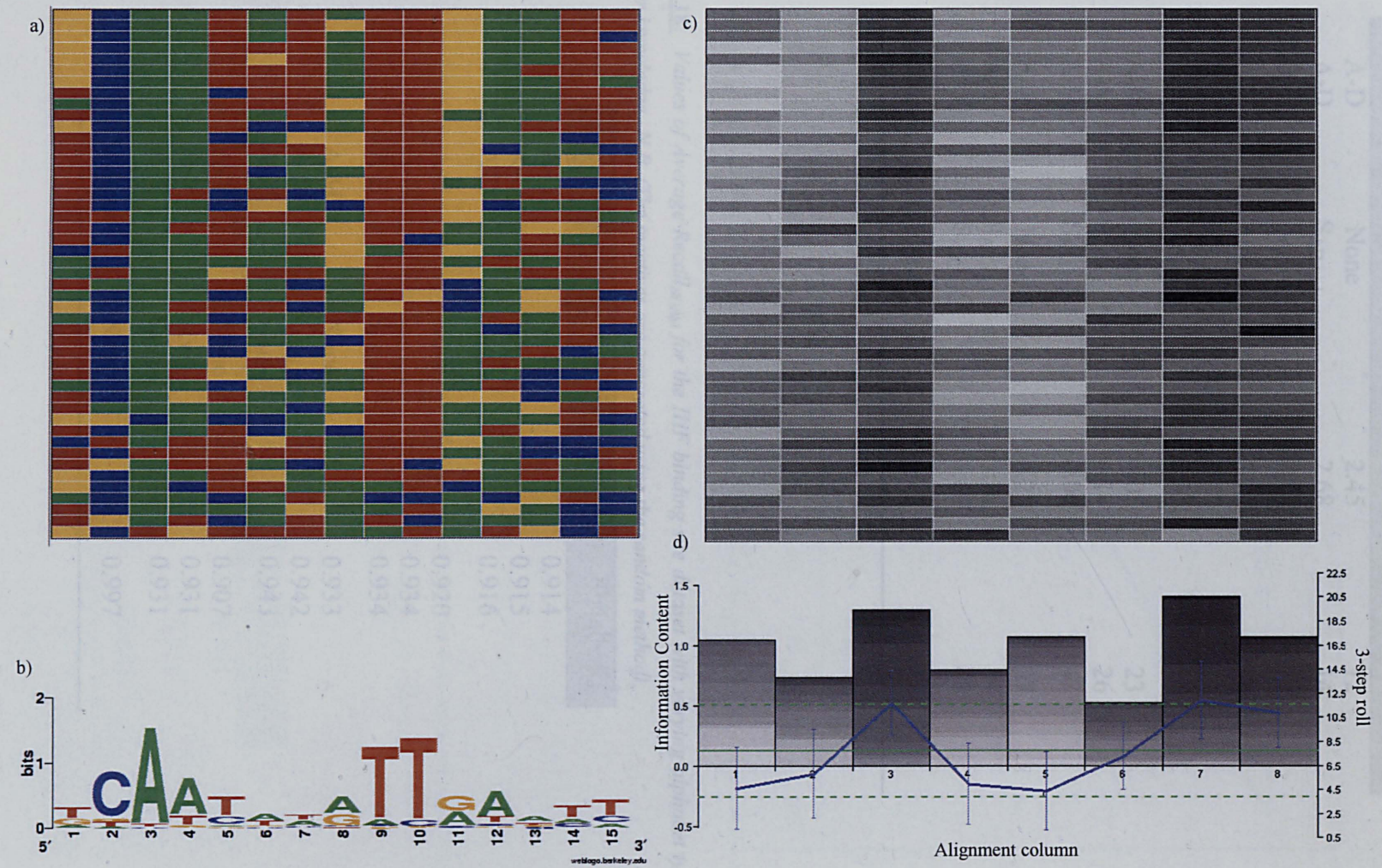


Table 9.17: LOO CV *scores for the IHF binding site dataset with varying alphabet type and prior knowledge.*

Alphabet	Prior knowledge	LOO CV Score	% Score reduction
A-D	None	2.45	12
A-D	$S_{AD,i \leftarrow j}$	2.68	14
A-D	$S_{AD,i=j}$	2.65	13
A-E	None	3.27	14
A-E	$S_{AE,i \leftarrow j}$	3.57	14
A-E	$S_{AE,i=j}$	3.33	18
A-M	None	4.05	23
A-M	$S_{AM,i \leftarrow j}$	4.14	26
A-M	$S_{AM,i=j}$	3.06	43
A-Y	None	4.28	37
A-Y	$S_{AY,i \leftarrow j}$	3.31	54
A-Y	$S_{AY,i=j}$	3.65	48
Sequence	None	9.12	11

Table 9.18: *Values of Average* $Recall_{NORM}$ *for the IHF binding site dataset with varying alphabet type and prior knowledge. N.B. (The inactives are generated using the random method).*

Alphabet	Prior knowledge	Average $Recall_{Norm}$
A-D	None	0.914
A-D	$S_{AD,i \leftarrow j}$	0.915
A-D	$S_{AD,i=j}$	0.916
A-E	None	0.920
A-E	$S_{AE,i \leftarrow j}$	0.934
A-E	$S_{AE,i=j}$	0.934
A-M	None	0.933
A-M	$S_{AM,i \leftarrow j}$	0.942
A-M	$S_{AM,i=j}$	0.943
A-Y	None	0.907
A-Y	$S_{AY,i \leftarrow j}$	0.931
A-Y	$S_{AY,i=j}$	0.931
Sequence	None	0.997

Figure 9.20: *Cumulative recall plot for the IHF binding site dataset with varying alphabet type and prior knowledge. A-D is dark blue, A-E is red, A-M is green, A-Y is light blue, sequence is orange, random recall is black and ideal recall is bright green. N.B. (The inactives are generated using the random method).*

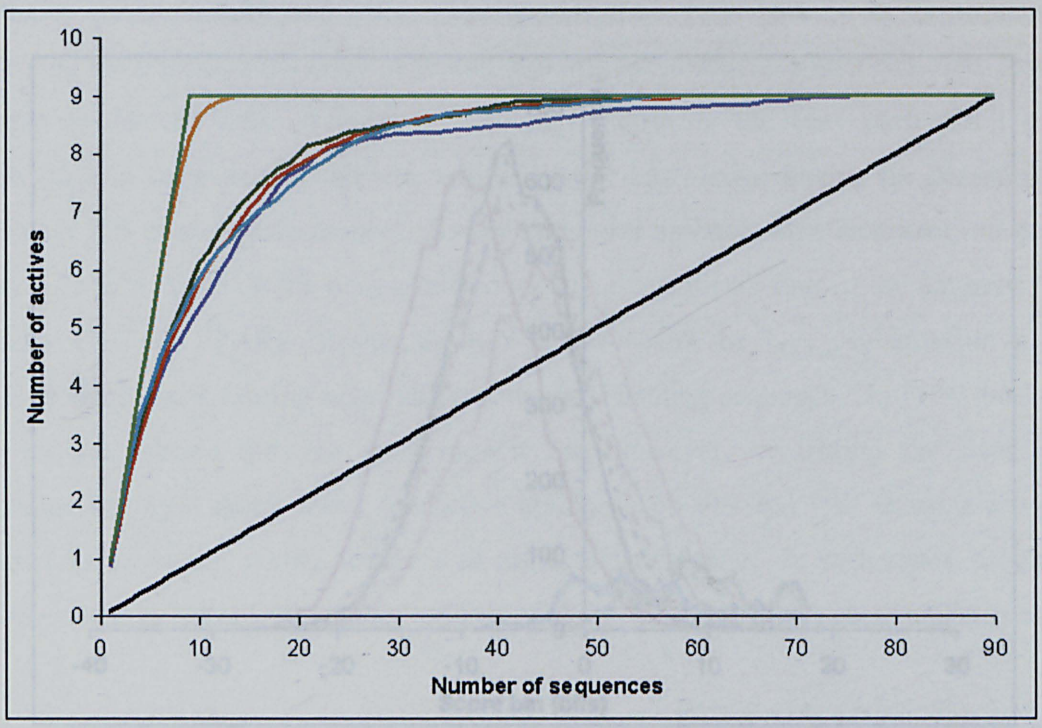


Table 9.19: *Exploring the effect different methods for generating the inactives have upon values of* $Recall_{NORM}$ *for the sequence and A-Y (with* $S_{AY,i \leftarrow j}$*)models for the IHF binding site dataset.*

Inactive Class	Sequence $Recall_{NORM}$	A-Y $Recall_{NORM}$
Random	0.997	0.931
Mono shuffling	0.981	0.918
Dinucleotide shuffling	0.930	0.885

Figure 9.21: *IHF score distributions of the sequence inactives (red), sequence actives (green), structure inactives (purple) and structure actives (light blue). The level of shading refers to the method used to generate the inactives: random generation (solid lines), mono shuffling (dashed lines) and doublet shuffling (lighter shading). The $S_{i \leftarrow j}$ substitution matrix is used for structure.*

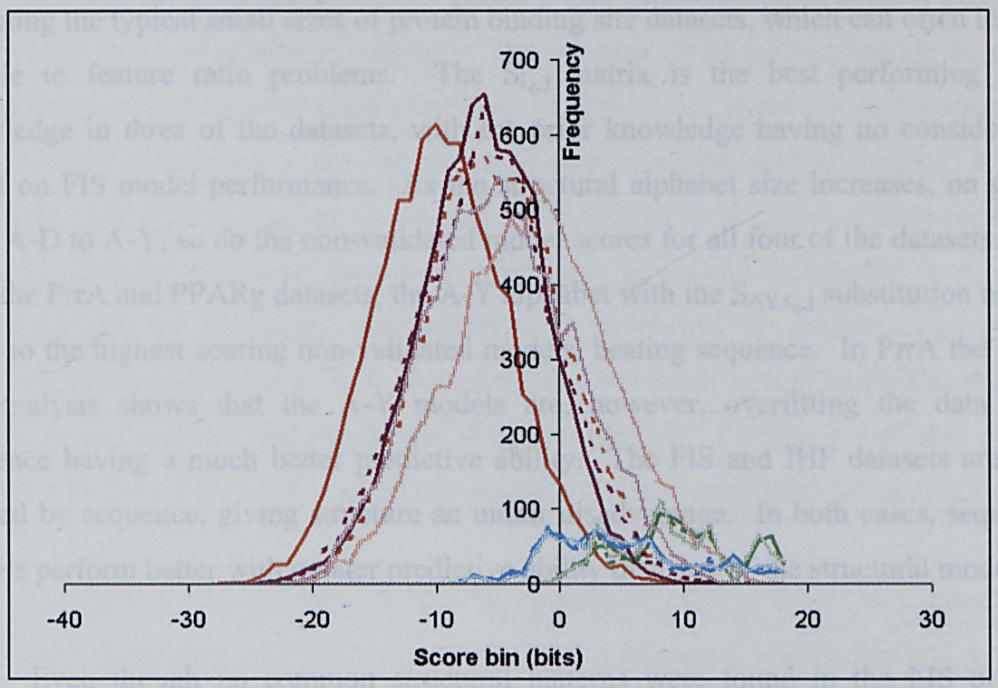


Table 9.20: *T-test analyses between active and inactive score distributions from Figure 9.21. DOF is the number of degrees of freedom. $\bar{x}_1 - \bar{x}_2$ is the mean separation between the two distributions. The $S_{i \leftarrow j}$ substitution matrix is used for structure.*

Alphabet	Type of inactives	DOF	T	$\bar{x}_1 - \bar{x}_2$
Sequence	Random	1163	111.88	19.02
Sequence	Mono	1182	87.20	14.74
Sequence	Doublet	1277	66.52	11.61
A-Y	Random	1054	54.03	11.01
A-Y	Mono	1088	53.10	10.86
A-Y	Doublet	1117	48.08	10.00

9.5. Conclusions

All the datasets have an average pairwise structural similarity greater than sequence (between 0.80 and 0.85). The dataset sizes vary from 38 to 91 sequences, reflecting the typical small sizes of protein binding site datasets, which can often lead to sample to feature ratio problems. The $S_{i \leftarrow j}$ matrix is the best performing prior knowledge in three of the datasets, with the prior knowledge having no considerable effect on FIS model performance. As the structural alphabet size increases, on going from A-D to A-Y, so do the non-validated model scores for all four of the datasets. For both the PrrA and PPARg datasets, the A-Y alphabet with the $S_{AY,i \leftarrow j}$ substitution matrix leads to the highest scoring non-validated models, beating sequence. In PrrA the LOO CV analysis shows that the A-Y models are, however, overfitting the data with sequence having a much better predictive ability. The FIS and IHF datasets are pre-aligned by sequence, giving structure an unfair disadvantage. In both cases, sequence models perform better with greater predictive ability than any of the structural models.

Even though no common structural patterns were found in the FIS or IHF binding sites, the structural models still had excellent recall ability and significant differences in the scores they assigned to the active and inactive test set sequences. In fact, for FIS, structure seems more stable in its predictions than sequence when going from the randomly generated inactives to the doublet shuffled inactives. This points to the previously mentioned weakness of the test set recall validation; random sequences will always be very different from those of the dataset despite mono- or di-nucleotide shuffling. In the IHF data, the A-Y models give a greater mean separation between the actives and inactives than sequence. However the significance of this separation is less due to greater variation within the scores of a distribution.

The structural alignment of the PrrA binding sites identified two important alignment positions (6 and 20) where 3-step roll tends to be greater than average. In the observed PPARg alignment a repeating low to high roll fingerprint is seen, with gaps never interrupting a sequence.

Chapter 10:
Conclusions and Future Research

The aim of this thesis was to develop and use tools that analyse how the structure of DNA varies with its function. The Octamer Database was used to describe the minimum energy structure and flexibility of DNA. An extension to the database was presented, calculating structural probabilities to describe DNA dynamics. It was discovered that a large number of octamer pairs that have identical or near identical minimum energy structures have very different structural tendencies. A Java application (Profile Manager) was successfully developed to analyse any special structural features of a single DNA sequence. The use of structural profiles to explore patterns across multiple sequences was then investigated. Finally, a tool that aligns sequences by their 3-step roll to obtain structural activity fingerprints was implemented.

This thesis will now conclude with discussion of five key topics: parameter correlations, flexibility and DNA dynamics, Profile Manager, hidden Markov models and architectural suppression.

10.1. Parameter correlations

When considering the degrees of freedom that describe the geometry between two base-pairs, the translations and rotations along each axis are correlated (that is shift and tilt for the x-axis, slide and roll for the y-axis, and rise and twist for the z-axis). This is because steric clashes that are caused by one movement can be minimised by changes in the other movement. The exact relationship between slide and roll was found to be somewhat more complicated than the analogous x-axis and z-axis correlations, since it is highly dependent on the central step type. Eight out of the ten step types have strong inverse slide-roll relationships as expected. However, the two guanine-pyrimidine steps have a wave-like slide-roll correlation. The structural reasons for this are unknown and need further investigation. Twist-roll plots that are identical in nature to the slide-roll plots were found, with twist strongly correlated to slide when considering the different central step types separately.

Another puzzling correlation found within this work was between the location preference and k^-_{twist} smoothed 30 promoter profiles. An attempt to confirm the associated parameter correlation was unsuccessful. The tri-nucleotide location preference was converted into an octamer descriptor by simply summing the six overlapping trimer values. Other methods for converting a tri-nucleotide descriptor into an octamer descriptor should be studied.

10.2. Flexibility and DNA dynamics

On average, increasing twist is the most favoured direction in flexibility. An octamer never appears to be highly rigid with respect to both decrease in roll and increase in twist, therefore there is always some degree of flexibility for widening the major groove (a common way by which proteins bind to DNA). A large number of octamer pairs that have identical or near identical minimum energy structures have very different structural tendencies. The probability of two identical octamers having the exact same conformation at one given time is low, reflecting the importance of dynamics in DNA structure. It can be more probable for two different octamers to have the same central step geometry than two identical octamers. This happens when an extremely flexible octamer has a minimum energy structure very close to that of a rigid octamer.

The single sequence profiles of two promoters highlighted frequent transitions in flexibility. This led to the hypothesis that sudden changes in flexibility may be an important promoter feature with flexible octamers putting stress upon the surrounding rigid octamers and presenting sites where the double helix can be easily unravelled for transcription initiation. The well-known TBP-TATA complex supports this importance of promoter flexibility and DNA dynamics. When multiple promoter sequences were considered, k^+_{twist} was the most distinguishing direction in flexibility in comparison to the octamer population. k^-_{twist} was found to be important at –100 to 0 relative to the transcription start site with a large transition becoming apparent in the TATA region after a smoothing window of 30 was applied to the promoter profile. Favourable

conformational changes in average promoter involve increases in twist with decreases having a high energetic penalty. Note that due to the transition in flexibility, the energy barrier for decreasing twist is suddenly reduced at the transcription start site.

10.3. Profile Manager

Profile Manager is a valuable visualisation tool for the analysis of DNA structure. This was clearly illustrated by the A-tract example with characteristic patterns in roll and minor groove width. The next stage in the development of Profile Manager is to gain user feedback and to decide whether it should remain as an application or be made available on the Internet.

10.4. Hidden Markov Models

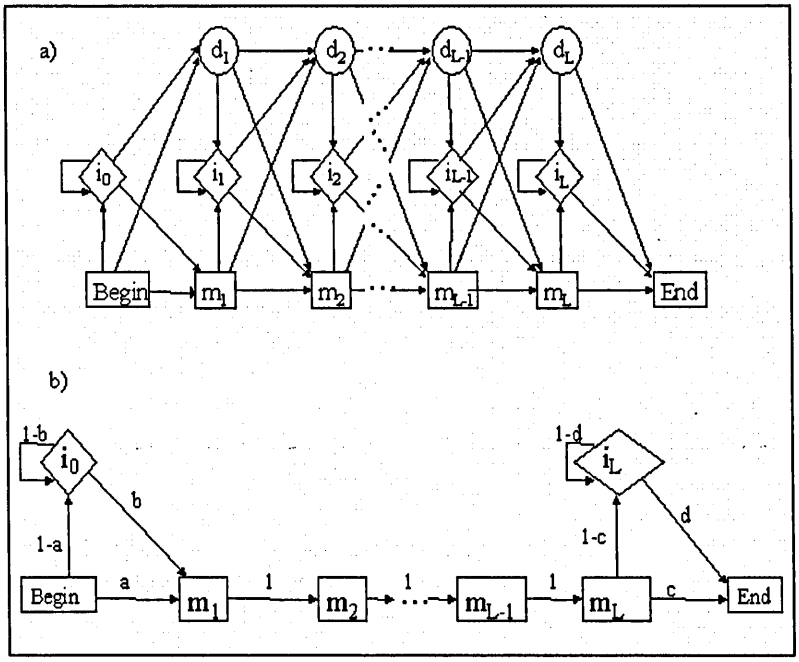
Structural 3-step roll alignments of DNA can now be successfully obtained from the novel HMM extension presented here. Structural alphabets of different sizes were explored and two subtly different substitution matrices for defining the inter-bin relationships were introduced. The $S_{i=j}$ matrix has a lower entropy than $S_{i\leftarrow j}$, meaning that it has a lower level of differentiation between the substitution probabilities and background probabilities. Indeed, it was found that $S_{i\leftarrow j}$ generally gave superior results to $S_{i=j}$. $S_{AY,i=j}$ sees the possibility that distant roll bins can be structurally equivalent, whereas $S_{AY,i\leftarrow j}$ strongly forbids distant bins to have equivalent structures. The A-Y alphabet seems to form the best models, but caution should be given to overfitting. Structural alignments of PrrA and PPARg were comparable in performance to sequence with useful insights into structure.

Future work should include generating the analogous 3-step twist alignment tool and ultimately to explore the possibility of performing a combined roll-twist alignment. Bistability of octamers could also be encoded into an HMM by sequence weights. Further HMM applications worth exploring are: promoter recognition, splice site detection and nucleosome wrapping.

10.5. Architectural suppression

The key to generating a successful HMM is adapting the model architecture and defining the allowed transitions in relation to the problem domain (Durbin et al., 1998). A gap-less alignment is required when studying DNA nucleosomal wrapping regions. For this purpose, the suppressed HMM architecture of Figure 10.1b should be used with the model length L being set to 147 (the known wrapping length). Note that in comparison to the traditional biological sequence HMM architecture (Figure 10.1a) a large number of transitions between states have been suppressed with only 4 unknowns (a, b, c and d) remaining in the transition matrix. The hidden element of the model is maintained, since different state paths that lead to the same observation sequence still exist, except when the sequence length, N, is equal to L.

Figure 10.1: *The Architectures. a)traditional, b)suppressed.*



Since the suppressed architecture is a highly constrained adaptation of the traditional it will have a much smaller solution search space. The above architecture can be implemented by partially freezing the traditional architecture and optimising emissions. A transition from state k to state l can be disabled by setting the maximum likelihood estimator a_{kl} to 0.

10.6. Concluding Remark

Even when a structural model does not have a greater predictive ability than a sequence model, it will still provide additional information about a structural mechanism that could never be gained from looking purely at the nucleotide sequence. This thesis has introduced novel tools for providing such additional information.

References

Acton, F.S. (1990). *Numerical methods that work.* Washington: The Mathematical Association of America.

Al-Lazikani, B., Sheinerman, F.B. & Honig, B. (2001). "Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases". *Proceedings of the National Academy of Sciences USA*, **98** (26), 14796-801.

Altschul, S.F. (1991). "Amino-acid substitution matrices from an information theoretic perspective". *Journal of Molecular Biology*, **219** (3), 555-565.

Altschul, S.F. & Erickson, B.W. (1985). "Significance of nucleotide-sequence alignments - a method for random sequence permutation that preserves dinucleotide and codon usage". *Molecular Biology And Evolution*, **2** (6), 526-538.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M.D., Roach, J., Oh, T., Ho, I.Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S.F., Clark, M.S., Edwards, Y.J., Doggett, N., Zharkikh, A., Tavtigian, S.V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y.H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D. & Brenner, S. (2002). "Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes". *Science*, **297** (5585), 1301-1310.

Baldi, P., Brunak, S., Chauvin, Y., Engelbrecht, J. & Krogh, A. (1995). "Periodic sequence patterns in human exons". In: Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T. & Wodak, S. (eds.), *Proceedings of the 3rd International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, CA. pp. 30-38. AAAI Press, Menlo Park, CA.

Baldi, P., Brunak, S., Chauvin, Y. & Krogh, A. (1996). "Naturally occurring nucleosome positioning signals in human exons and introns". *Journal of Molecular Biology*, **263** (4), 503-510.

Baldi, P., Chauvin, Y., Hunkapiller, T. & McClure, M.A. (1994). "Hidden Markov models of biological primary sequence information". *Proceedings of the National Academy of Sciences USA*, **91** (3), 1059-1063.

Barash, Y., Elidan, G., Freidman, N. & Kaplan, T. (2003). "Modeling dependencies in protein-DNA binding sites". *Proceedings of the 7th annual international conference on Computational Molecular Biology*, pp. 28-37.

Barrett, C., Hughey, R. & Karplus, K. (1997). "Scoring hidden Markov models". *Computer Applications In The Biosciences*, **13** (2), 191-199.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. & Sonnhammer, E.L.L. (2000). "The Pfam protein families database". *Nucleic Acids Research*, **28** (1), 263-266.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C. & Eddy, S.R. (2004). "The Pfam protein families database". *Nucleic Acids Research*, **32**, D138-D141.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. & Haussler, D. (2004). "Ultraconserved elements in the human genome". *Science*, **304** (5675), 1321-1325.

Bengio, Y. (1999). "Markovian models for sequential data". *Neural Computing Surveys*, **2**, 129-162.

Benson, D.A., Boguski, M.S., Lipman, D.J. & Ostell, J. (1997). "GenBank". *Nucleic Acids Research*, **25** (1), 1-6.
Berger, J. & Moller, D.E. (2002). "The mechanism of action of PPARs". *Annual Reviews Medicine*, **53**, 409-435.
Bilmes, J. (2002). *What HMMs can do*. University of Washington. (UW Electrical Energineering Technical Report Report No. UWEETR-2002-0003).
Borodovsky, M. & McIninch, J. (1993). "Genmark - parallel gene recognition for both DNA strands". *Computers and Chemistry*, **17** (2), 123-133.
Borse, G.J. (1997). *Numerical methods with MATLAB*. Boson, MA: PWS Publishing Company.
Branden, C. & Tooze, J. (1991). *Introduction to protein structure*. London: Garland Publishing Inc.
Brukner, I., Sanchez, R., Suck, D. & Pongor, S. (1995). "Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides". *European Molecular Biology Organisation Journal*, **14** (8), 1812-8.
Bucher, P., Karplus, K., Moeri, N. & Hofmann, K. (1996). "A flexible motif search technique based on generalized profiles". *Computers and Chemistry*, **20** (1), 3-23.
Bussard, A.E. (2005) "A scientific revolution?". *European Molecular Biology Organisation Reports*, **6**(8), 691-694.
Butina, D. (1999). "Unsupervised database clustering based on Daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets". *Journal of Chemical Information Computer Science*, **39**, 747-750.
Butler, J.E. & Kadonaga, J.T. (2002). "The RNA polymerase II core promoter: a key component in the regulation of gene expression". *Genes and Development*, **16** (20), 2583-92.
Bystroff, C. & Baker, D. (1998). "Prediction of local structure in proteins using a library of sequence-structure motifs". *Journal of Molecular Biology*, **281** (3), 565-577.
Bystroff, C., Thorsson, V. & Baker, D. (2000). "HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins". *Journal of Molecular Biology*, **301** (1), 173-190.
Calladine, C.R. & Drew, H.R. (2002). *Understanding DNA*. London: Academic Press.
Chechetkin, V.R. (2003). "Block structure and stability of the genetic code". *Journal of Theoretical Biology*, **222** (2), 177-188.
Chen, X.J. & Butow, R.A. (2005). "The organisation and inheritance of the mitochondrial genome". *Nature Reviews*, **6**, 815-825.
Churchill, G.A. (1989). "Stochastic models for heterogeneous DNA sequences". *Bulletin of Mathematical Biology*, **51** (1), 79-94.
Churchill, G.A. (1992). "Hidden Markov chains and the analysis of genome structure". *Computers and Chemistry*, **16** (2), 107-115.
Claverie, J.M. (1992). "Sequence signals - artifact or reality". *Computers and Chemistry*, **16** (2), 89-91.
Collins, F.S., Green, E.D., Guttmacher, A.E. & Guyer, M.S. (2003b). "A vision for the future of genomics research". *Nature*, **422** (6934), 835-847.
Collins, F.S., Morgan, M. & Patrinos, A. (2003a). "The human genome project: lessons from large-scale biology". *Science*, **300** (5617), 286-290.
Coornaert, D., Vissers, S., Andre, B. & Grenson, M. (1992). "The UGA43 negative regulatory gene of Saccharomyces cerevisiae contains both a GATA-1 type zinc finger and a putative leucine zipper". *Current Genetics*, **21** (4-5), 301-307.
Coward, E. (1999). "Shufflers: shuffling sequences while conserving the k-let counts". *Bioinformatics*, **15** (12), 1058-1059.

Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. (2004). "WebLogo: a sequence logo generator". *Genome Research*, **14** (6), 1188-1190.

Daar, A.S., Bhatt, A., Court, E. & Singer, P.A. (2004). "Stem cell research and transplantation: science leading ethics". *Transplant Proceedings*, **36** (8), 2504-2506.

Daly, F., Hand, D.J., Jones, M.C., Lunn, A.D. & McConway, K.J. (1995). *Elements of statistics.* Wokingham, England: Addison-Wesley Publishing Company.

Davis, N.A., Majee, S.S. & Kahn, J.D. (1999). "TATA box DNA deformation with and without the TATA box-binding protein". *Journal of Molecular Biology*, **291** (2), 249-65.

Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978). "A model of evolutionary change in proteins". *Atlas of Protein Sequence and Structure*, **5** (3), 345-352.

Di Francesco, V., Munson, P.J. & Garnier, J. (1999). "FORESST: fold recognition from secondary structure predictions of proteins". *Bioinformatics*, **15** (2), 131-140.

Diekmann, S. (1989). "Definitions and nomenclature of nucleic-acid structure parameters". *Journal of Molecular Biology*, **205** (4), 787-791.

DiFrancesco, V., Garnier, J. & Munson, P.J. (1997). "Protein topology recognition from secondary structure sequences: application of the hidden Markov models to the alpha class proteins". *Journal of Molecular Biology*, **267** (2), 446-463.

Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge University Press, Cambridge, UK.

Eddy, S.R. (1995). "Multiple alignment using hidden Markov models". *Proceedings of 3rd Internatioal Conference Intelligent Systems for Molecular Biology*, Menlo Park, CA. pp. 114-120. AAAI Press, Menlo Park, CA.

Eddy, S.R. (1996). "Hidden Markov models". *Current Opinion In Structural Biology*, **6** (3), 361-365.

Eddy, S.R. (1998). "Profile hidden Markov models". *Bioinformatics*, **14** (9), 755-763.

Eddy, S.R., Mitchison, G. & Durbin, R. (1995). "Maximum discrimination hidden Markov models of sequence consensus". *Journal of Computational Biology*, **2**, 9-23.

El-Hassan, M.A. & Calladine, C.R. (1995). "The assessment of the geometry of dinucleotide steps in double-helical DNA - a new local calculation scheme". *Journal of Molecular Biology*, **251** (5), 648-664.

El-Hassan, M.A. & Calladine, C.R. (1996). "Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA". *Journal of Molecular Biology*, **259** (1), 95-103.

Foster, M.W. & Sharp, R.R. (2000). "Genetic research and culturally specific risks: one size does not fit all". *Trends in Genetics*, **16** (2), 93-95.

Frazier, M.E., Johnson, G.M., Thomassen, D.G., Oliver, C.E. & Patrinos, A. (2003). "Realizing the potential of the genome revolution: the genomes to life program". *Science*, **300** (5617), 290-293.

Frommlet, F., Futschik, A. & Bogdan, M. (2004). "On the significance of sequence alignments when using multiple scoring matrices". *Bioinformatics*, **20** (6), 881-887.

Gander, W. & Gautschi, W. (2000). "Adapted quadrature-revisited". *BIT*, **40**, 84-101.

Gardiner, E.J., Hunter, C.A., Lu, X.J. & Willett, P. (2004). "A structural similarity analysis of double-helical DNA". *Journal of Molecular Biology*, **343** (4), 879-89.

Gardiner, E.J., Hunter, C.A., Packer, M.J., Palmer, D.S. & Willett, P. (2003). "Sequence-dependent DNA structure: a database of octamer structural parameters". *Journal of Molecular Biology*, **332** (5), 1025-1035.

German, M.S., Wang, J., Chadwick, R.B. & Rutter, W.J. (1992). "Synergistic activation of the insulin gene by a LIM-homeo domain protein and a basic helix-loop-helix protein: building a functional insulin minienhancer complex". *Genes Development*, **6** (11), 2165-2176.

Gerstein, M., Sonnhammer, E.L.L. & Chothia, C. (1994). "Volume changes in protein evolution". *Journal of Molecular Biology*, **236** (4), 1067-1078.

Glover, J.N. & Harrison, S.C. (1995). "Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA". *Nature*, **373** (6511), 257-261.

Gonnet, G.H., Cohen, M.A. & Benner, S.A. (1994). "Analysis of amino-acid substitution during divergent evolution - the 400 by 400 dipeptide substitution matrix". *Biochemical And Biophysical Research Communications*, **199** (2), 489-496.

Goodsell, D.S. & Dickerson, R.E. (1994). "Bending and curvature calculations in B-DNA". *Nucleic Acids Research*, **22** (24), 5497-503.

Gorin, A.A., Zhurkin, V.B. & Olson, W.K. (1995). "B-DNA twisting correlates with base-pair morphology". *Journal of Molecular Biology*, **247** (1), 34-48.

Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure". *Journal of Molecular Biology*, **313** (4), 903-919.

Grimmett, G. & Stirzaker, D. (2001). *Probability and random processes.* New York: Oxford University Press Inc.

Hai, T.W., Liu, F., Coukos, W.J. & Green, M.R. (1989). "Transcription factor ATF cDNA clones: an extensive family of leucine zipper proteins able to selectively form DNA-binding heterodimers". *Genes Development*, **3**, 2083-2090.

Hasan, S. (2003). *Prediction and analysis of nucleosome positioning in genomic sequences*. Wolfson College, University of Cambridge.

Haussler, D., Krogh, A., Mian, I.S. & Sjolander, K. (1993). "Protein modeling using hidden Markov models: analysis of globins". *Proc. of 26th Annual Hawaii International Conference on System Sciences*, Vol. 1, Los Alamitos, California. pp. 792-802. IEEE Computer Society Press, Los Alamitos, California.

Hawkins, D.M. (2004). "The problem of overfitting". *Journal of Chemical Information Computer Science*, **44** (1), 1-12.

Hawkins, D.M., Basak, S.C. & Mills, D. (2003). "Assessing model fit by cross-validation". *Journal of Chemical Information Computer Science*, **43** (2), 579-86.

Hawkins, D.M., Basak, S.C. & Shi, X. (2001). "QSAR with few compounds and many features". *Journal of Chemical Information Computer Science*, **41** (3), 663-70.

Helene, C. (1998). "DNA recognition: reading the minor groove". *Nature*, **391**, 436-8.

Henikoff, S. & Henikoff, J.G. (1992). "Amino-acid substitution matrices from protein blocks". *Proceedings Of The National Academy Of Sciences USA*, **89** (22), 10915-10919.

Holbrook, P.G., Geetha, V., Beaven, M.A. & Munson, P.J. (1999). "Recognizing the pleckstrin homology domain fold in mammalian phospholipase D using hidden Markov models". *Febs Letters*, **448** (2-3), 269-272.

Horton, H.R., Moran, L.A., Ochs, R.S., Rawn, J.D. & Scrimgeour, K.G. (2002). *Principles of biochemistry.* Prentice-Hall Inc.

Hourai, Y., Akutsu, T. & Akiyama, Y. (2004). "Optimizing substitution matrices by separating score distributions". *Bioinformatics*, **20** (6), 863-873.

Hsu, T., Gogos, J.A., Kirsh, S.A. & Kafatos, F.C. (1992). "Multiple zinc finger forms resulting from developmentally regulated alternative splicing of a transcription factor gene". *Science*, **257** (5078), 1946-1950.

Hunter, C.A. (1993). "Sequence-dependent DNA structure - the role of base stacking interactions". *Journal of Molecular Biology*, **230** (3), 1025-1054.

Hunter, C.A. & Lu, X.J. (1997). "DNA base-stacking interactions: a comparison of theoretical calculations with oligonucleotide x-ray crystal structures". *Journal of Molecular Biology*, **265** (5), 603-619.

Johnston, L. & Hagan, P. (2003). "Eat up your carrots and beat measles". *Daily Express*, 17th May, 33.

Johnston, M. & Stormo, G.D. (2003). "Evolution: heirlooms in the attic". *Science*, **302** (5647), 997-999.

Juo, Z.S., Chiu, T.K., Leiberman, P.M., Baikalov, I., Berk, A.J. & Dickerson, R.E. (1996). "How proteins recognize the TATA box". *Journal of Molecular Biology*, **261** (2), 239-54.

Kandel, D., Matias, Y., Unger, R. & Winkler, P. (1996). "Shuffling biological sequences". *Discrete Applied Mathematics*, **71** (1-3), 171-185.

Karchin, R. & Hughey, R. (1998). "Weighting hidden Markov models for maximum discrimination". *Bioinformatics*, **14** (9), 772-782.

Karplus, K., Barrett, C. & Hughey, R. (1998). "Hidden Markov models for detecting remote protein homologies". *Bioinformatics*, **14** (10), 846-856.

Kasif, S. & Delcher, A.L. (1998). "Modeling biological data and structure with probabilistic networks." In: Salzberg, S.L., Searls, D.B. & Kasif, S. (eds.), *Computational Methods in Molecular Biology*, pp. 335-352. Elsevier Science.

Kato, R., Noguchi, H., Honda, H. & Kobayashi, T. (2003). "Hidden Markov model-based approach as the first screening of binding peptides that interact with MHC class II molecules". *Enzyme and Microbial Technology*, **33**, 472-481.

Kielkopf, C.L., White, S., Szewczyk, J.W., Turner, J.M., Baird, E.E., Dervan, P.B. & Rees, D.C. (1998). "A structural basis for recognition of A.T and T.A base pairs in the minor groove of B-DNA". *Science*, **282** (5386), 111-115.

Kikuchi, N., Kwon, Y.D., Gotoh, M. & Narimatsu, H. (2003). "Comparison of glycosyltransferase families using the profile hidden Markov model". *Biochemical Biophysics Research Communications*, **310** (2), 574-9.

Knight, R.D., Freeland, S.J. & Landweber, L.F. (1999). "Selection, history and chemistry: the three faces of the genetic code". *Trends in Biochemical Science*, **24** (6), 241-247.

Koski, A. (1996). "Modelling ECG signals with hidden Markov models". *Artifical Intelligence in Medicine*, **8** (5), 453-71.

Koudelka, G.B., Harbury, P., Harrison, S.C. & Ptashne, M. (1988). "DNA twisting and the affinity of bacteriophage-434 operator for bacteriophage-434 repressor". *Proceedings Of The National Academy Of Sciences USA*, **85** (13), 4633-4637.

Krogh, A. (1994c). "Hidden Markov models for labeled sequences". *Proceedings of 12th International Conference on Pattern Recognition*, Los Alamitos, California. pp. 140-144. IEEE Computer Society Press, Los Alamitos, California.

Krogh, A. (1997). "Two methods for improving performance of an HMM and their application for gene finding". In: Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. & Valencia, A. (eds.), *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, CA. pp. 179-186. AAAI Press, Menlo Park, CA.

Krogh, A. (1998). "An introduction to hidden Markov models for biological sequences". In: Salzberg, S.L., Searls, D.B. & Kasif, S. (eds.), *Computational Methods in Molecular Biology*, pp. 45-63. Elsevier Science.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K. & Haussler, D. (1994a). "Hidden Markov models in computational biology: applications to protein modeling". *Journal of Molecular Biology*, **235** (5), 1501-31.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes". *Journal of Molecular Biology*, **305** (3), 567-80.

Krogh, A., Mian, I.S. & Haussler, D. (1994b). "A hidden Markov model that finds genes in E. coli DNA". *Nucleic Acids Research*, **22** (22), 4768-78.

Kutach, A.K. & Kadonaga, J.T. (2000). "The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters". *Molecular Cell Biology*, **20** (13), 4754-4764.

Laguri, C., Phillips-Jones, M.K. & Williamson, M.P. (2003). "Solution structure and DNA binding of the effector domain from the global regulator PrrA (RegA) from Rhodobacter sphaeroides: insights into DNA binding specificity". *Nucleic Acids Research*, **31** (23), 6778-6787.

Lean, G. (2005). "Revealed: health fears over secret study into GM food". *The Independent*, 22nd May,

Lesk, A.M. (2003). *Introduction to bioinformatics.* Oxford: Oxford Univerity Press.

Levitt, M. (1983). "Computer simulation of DNA double-helix dynamics". *Cold Spring Harbour Symposia on Quantitative Biology*, **47 Pt 1**, 251-62.

Lisker, R. (2003). "Ethical and legal issues in therapeutic cloning and the study of stem cells". *Archives of Medical Research*, **34** (6), 607-611.

Liu, Q., Zhu, Y.S., Wang, B.H. & Li, Y.X. (2003). "A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins". *Computational Biology and Chemistry*, **27** (1), 69-76.

Lu, X.J., ElHassan, M.A. & Hunter, C.A. (1997a). "Structure and conformation of helical nucleic acids: analysis program (SCHNAaP)". *Journal of Molecular Biology*, **273** (3), 668-680.

Lu, X.J., ElHassan, M.A. & Hunter, C.A. (1997b). "Structure and conformation of helical nucleic acids: rebuilding program (SCHNArP)". *Journal of Molecular Biology*, **273** (3), 681-691.

Lu, X.J. & Olson, W.K. (2003). "3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures". *Nucleic Acids Research*, **31** (17), 5108-5121.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F. & Richmond, T.J. (1997). "Crystal structure of the nucleosome core particle at 2.8 A resolution". *Nature*, **389** (6648), 251-260.

Lukashin, A.V. & Borodovsky, M. (1998). "GeneMark.hmm: new solutions for gene finding". *Nucleic Acids Research*, **26** (4), 1107-15.

Mager, J. & Bartolomei, M.S. (2000). "Strategies for dissecting epigenetic mechanisms in the mouse", *Nature Genetics*, **37**(11), 1194-1200.

Mandel, T. (1997). *The elements of user interface design.* John Wiley & Sons, Inc.

Martinez, A. (1999). "Face image retrieval using HMMs". *Proceedings of IEEE Workshop on Content-Based Access of Images and Video Libraries*, pp. 25-39.

McKenzie, R.W. & Brennan, M.D. (1996). "The two small introns of the Drosophila affinidisjuncta Adh gene are required for normal transcription". *Nucleic Acids Research*, **24** (18), 3635-3642.

Melnikoff, S.J., Quigley, S.F. & Russell, M.J. (2002). "Speech recognition on an FPA using discrete and continuous hidden Markov models". *Field-Programmable Logic And Applications, Proceedings*, Vol. 2438, pp. 202-211.

Miller, J.C. & Miller, J.N. (1994). *Statistics for analytical chemistry.* Chichester, W. Sussex: Ellis Horwood Limited.

Needleman, S.B. & Wuncsh, C.D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology*, **48**, 443-453.

Nikolov, D.B. & Burley, S.K. (1997). "RNA polymerase II transcription initiation: a structural view". *Proceedings of the National Academy of Sciences USA*, **94** (1), 15-22.

Nippert, I. (2002). "The pros and cons of human therapeutic cloning in the public debate". *Journal of Biotechnology*, **98** (1), 53-60.

Nobrega, M.A., Zhu, Y., Plajzer-Frick, I., Afzal, V. & Rubin, E.M. (2004). "Megabase deletions of gene deserts result in viable mice". *Nature*, **431** (7011), 988-993.

Norris, J.R. (1999). *Markov chains.* Cambridge: Cambridge University Press Inc.

Olson, M.V. (1995). "A time to sequence". *Science*, **270** (5235), 394-396.

Olson, M.V. (2002). "The human genome project: a player's perspective". *Journal of Molecular Biology*, **319** (4), 931-942.

Packer, M.J., Dauncey, M.P. & Hunter, C.A. (2000a). "Sequence-dependent DNA structure: dinucleotide conformational maps". *Journal of Molecular Biology*, **295** (1), 71-83.

Packer, M.J., Dauncey, M.P. & Hunter, C.A. (2000b). "Sequence-dependent DNA structure: tetranucleotide conformational maps". *Journal of Molecular Biology*, **295** (1), 85-103.

Packer, M.J. & Hunter, C.A. (1998). "Sequence-dependent DNA structure: the role of the sugar- phosphate backbone". *Journal of Molecular Biology*, **280** (3), 407-420.

Packer, M.J. & Hunter, C.A. (2001). "Sequence-structure relationships in DNA oligomers: a computational approach". *Journal Of The American Chemical Society*, **123** (30), 7399-7406.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). "Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods". *Journal of Molecular Biology*, **284** (4), 1201-10.

Patterton, H.G. & Graves, S. (2000). "DNAssist: the integrated editing and analysis of molecular biology sequences in windows". *Bioinformatics*, **16** (7), 652-3.

Pearson, H. (2003). "DNA: beyond the double helix". *Nature*, **421** (6921), 310-312.

Pedersen, A.G., Baldi, P., Chauvin, Y. & Brunak, S. (1998). "DNA structure in human RNA polymerase II promoters". *Journal of Molecular Biology*, **281** (4), 663-73.

Pedersen, A.G., Baldi, P., Chauvin, Y. & Brunak, S. (1999). "The biology of eukaryotic promoter prediction--a review". *Computers and Chemistry*, **23** (3-4), 191-207.

Petersen, L., Larsen, T.S., Ussery, D.W., On, S.L. & Krogh, A. (2003). "RpoD promoters in Campylobacter jejuni exhibit a strong periodic signal instead of a -35 box". *Journal of Molecular Biology*, **326** (5), 1361-72.

Petschek, J.P., Scheckelhoff, M.R., Mermer, M.J. & Vaughn, J.C. (1997). "RNA editing and alternative splicing generate mRNA transcript diversity from the Drosophila 4f-rnp locus". *Gene*, **204**, 267-276.

Prlic, A., Domingues, F.S. & Sippl, M.J. (2000). "Structure-derived substitution matrices for alignment of distantly related sequences". *Protein Engineering*, **13** (8), 545-550.

Ptashne, M. (1987). *A genetic switch.* Oxford: Blackwell Scientific.
Purves, W.K., Orians, G.H., Heller, H.C. & Sadava, D. (1997). *Life, the science of biology.* Sinauer Associates Inc.
Rabiner, L.R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". *Proceedings Of The IEEE*, **77** (2), 257-286.
Randic, M. & Vracko, M. (2000a). "On the similarity of DNA primary sequences". *Journal of Chemical Information Computer Science*, **40** (3), 599-606.
Randic, M., Vracko, M., Nandy, A. & Basak, S.C. (2000b). "On 3-D graphical representation of DNA primary sequences and their numerical characterization". *Journal of Chemical Information Computer Science*, **40** (5), 1235-44.
Rasmussen, T.K. & Krink, T. (2003). "Improved hidden Markov model training for multiple sequence alignment by a particle swarm optimization-evolutionary algorithm hybrid". *Biosystems*, **72** (1-2), 5-17.
Raven, P.H. & Johnson, G.B. (1996). *Biology.* London: Wm. C. Brown Publishers.
Rebar, E.J., Huang, Y., Hickey, R., Nath, A.K., Meoli, D., Nath, S., Chen, B., Xu, L., Liang, Y., Jamieson, A.C., Zhang, L., Spratt, S.K., Case, C.C., Wolffe, A. & Giordano, F.J. (2002). "Induction of angiogenesis in a mouse model using engineered transcription factors". *Natural Medicines*, **8** (12), 1427-1432.
Rice, P.A., Yang, S.W., Mizuuchi, K. & Nash, H.A. (1996). "Crystal structure of an IHF-DNA complex: a protein-induced DNA u-turn". *Cell*, **87** (7), 1295-1306.
Rigoll, G., Kosmala, A., Rottland, J. & Neukirchen, C. (1996). "A comparison between continuous and discrete density hidden Markov models for cursive handwriting recognition". *International Conference on Pattern Recognition*, Vol. 2, Vienna. pp. 205-209. Vienna.
Saleh-Lakha, S. & Glick, B.R. (2005). "Is the battle over genetically modified foods finally over?" *Biotechnology Advances*, **23** (2), 93-96.
Salton, G. & McGill, M. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill.
Satchwell, S.C., Drew, H.R. & Travers, A.A. (1986). "Sequence periodicities in chicken nucleosome core DNA". *Journal of Molecular Biology*, **191** (4), 659-75.
Schneider, T.D. & Stephens, R.M. (1990). "Sequence logos - a new way to display consensus sequences". *Nucleic Acids Research*, **18** (20), 6097-6100.
Schultz, S.C., Shields, G.C. & Steitz, T.A. (1991). "Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees". *Science*, **253** (5023), 1001-1007.
Sedgewick, R. (1988). *Algorithms.* Addison-Wesley.
Shaner, M.C., Blair, I.M. & Schneider, T.D. (1993). "Sequence logos: a powerful, yet simple, tool". In: Mudge, T.N., Milutinovic, V. & Hunter, L. (eds.), *Proceedings of 26th Annual Hawaii International Conference on System Sciences*, Vol. 1, Hawaii. pp. 813-821. Hawaii.
Shao, J. (1993). "Linear model selection by cross-validation". *Journal of American Statistics Association*, **88** (422), 486-494.
Shaw-Lee, R., Lissemore, J.L., Sullivan, D.T. & Tolan, D.R. (1992). "Alternative splicing of fructose 1,6-bisphosphate aldolase transcripts in Drosophila melanogaster predicts three isozymes". *Journal of Biological Chemistry*, **267** (6), 3959-3967.
Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S. & Haussler, D. (1996). "Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology". *Computational Applications in Biosciences*, **12** (4), 327-45.

Sonnhammer, E.L., Von Heijne, G. & Krogh, A. (1998). "A hidden Markov model for predicting transmembrane helices in protein sequences". In: Glasgow, J., Littlejohn, T., Major, F., Lathrop, R., Sankoff, D. & Sensen, C. (eds.), *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, CA. pp. 175-182. AAAI Press, Menlo Park, CA.

Stefl, R., Wu, H., Ravindranathan, S., Sklenar, V. & Feigon, J. (2004). "DNA A-tract bending in three dimensions: solving the dA4T4 vs. dT4A4 conundrum". *Proceedings of the National Academy of Sciences USA*, **101** (5), 1177-1182.

Steiner, E. (2000). *The chemistry maths book.* Oxford: Oxford University Press.

Stultz, C.M., White, J.V. & Smith, T.F. (1993). "Structural analysis based on state-space modeling". *Protein Science*, **2** (3), 305-314.

Swanson, H.I. & Yang, J.H. (1999). "Specificity of DNA binding of the c-Myc/Max and ARNT/ARNT dimers at the CACGTG recognition site". *Nucleic Acids Research*, **27** (15), 3205-3212.

Teodorescu, O., Galor, T., Pillardy, J. & Elber, R. (2004). "Enriching the sequence substitution matrix by structural information". *Proteins-Structure Function And Genetics*, **54** (1), 41-48.

Thayer, K.M. & Beveridge, D.L. (2002). "Hidden Markov models from molecular dynamics simulations on DNA". *Proceedings of the National Academy of Sciences USA*, **99** (13), 8642-7.

Thomas, P.D., Kejariwal, A., Campbell, M.J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S., Vandergriff, J.A. & Doremieux, O. (2003). "PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification". *Nucleic Acids Research*, **31** (1), 334-41.

Tiffany, H.L., Handen, J.S. & Rosenberg, H.F. (1996). "Enhanced expression of the eosinophil-derived neurotoxin ribonuclease (RNS2) gene requires interaction between the promoter and intron". *Journal of Biological Chemistry*, **271** (21), 12387-12393.

Travers, A. (1997). "DNA-protein interactions: IHF - the master bender". *Current Biology*, **7** (4), R252-R254.

Tusnady, G.E. & Simon, I. (1998). "Principles governing amino acid composition of integral membrane proteins: application to topology prediction". *Journal of Molecular Biology*, **283** (2), 489-506.

Ussery, D., Larsen, T.S., Wilkes, K.T., Friis, C., Worning, P., Krogh, A. & Brunak, S. (2001). "Genome organisation and chromatin structure in Escherichia coli". *Biochimie*, **83** (2), 201-212.

vanUlsen, P., Hillebrand, M., Zulianello, L., vandePutte, P. & Goosen, N. (1997). "The integration host factor-DNA complex upstream of the early promoter of bacteriophage Mu is functionally symmetric". *Journal Of Bacteriology*, **179** (9), 3073-3075.

Vilim, R.B., Cunningham, R.M., Lu, B., Kheradpour, P. & Stevens, F.J. (2004). "Fold-specific substitution matrices for protein classification". *Bioinformatics*, **20** (6), 847-853.

Vlahovicek, K., Kajan, L. & Pongor, S. (2003). "DNA analysis servers: plot.it, bend.it, model.it and IS". *Nucleic Acids Research*, **31** (13), 3686-7.

Vogt, G., Etzold, T. & Argos, P. (1995). "An assessment of amino-acid exchange matrices in aligning protein sequences - the twilight zone revisited". *Journal of Molecular Biology*, **249** (4), 816-831.

Wallhoff, F., Eickeler, S. & Rigoll, G. (2001). "A comparison of discrete and continuous output modelling techniques for a pseudo-2D hidden Markov model face recognition system". In: Thessaloniki (ed.), *IEEE Int. Conference on Image Processing*, Vol. October 2001, Greece. Greece.

Watson, J.D. & Crick, F.H. (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid". *Nature*, **171** (4356), 737-738.

Weinschenk, S., Jamar, P. & Yeo, S.C. (1997). *GUI design essentials for Windows 95, Windows 3.1 World Wide Web.* John Wiley & Sons, Inc.

Weiss, N. (1995). *Introductory statistics.* New York: Addison-Wesley Publishing Company Inc.

White, J.V., Stultz, C.M. & Smith, T.F. (1994). "Protein classification by stochastic modeling and optimal filtering of amino-acid sequences". *Mathematical Bioscience*, **119** (1), 35-75.

White, S., Szewczyk, J.W., Turner, J.M., Baird, E.E. & Dervan, P.B. (1998). "Recognition of the four Watson-Crick base pairs in the DNA minor groove by synthetic ligands". *Nature*, **391** (6666), 468-471.

Widom, J. (1997). "Chromosome structure and gene regulation". *Physica A*, **244**, 497-509.

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S. & Urbach, S. (2001). "The TRANSFAC system on gene expression regulation". *Nucleic Acids Research*, **29** (1), 281-283.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y.J., Cooke, J.E. & Elgar, G. (2005). "Highly conserved non-coding sequences are associated with vertebrate development". *PLoS Biol*, **3** (1), 116-130.

Wu, S. & Gu, X. (2002). "Random shuffling permutations of nucleotides". *The 6th world multiconference on Systematics, Cybernetics and Informatics*, Vol. 13, pp. 308-313.

Yanagi, K., Prive, G.G. & Dickerson, R.E. (1991). "Analysis of local helix geometry in 3 B-DNA decamers and 8 dodecamers". *Journal of Molecular Biology*, **217** (1), 201-214.

Yin, M. & Wang, J. (2001). "Effective hidden Markov models for detecting splicing junction sites in DNA sequences". *Information Sciences*, **139**, 139-163.

Yu, Y.K., Wootton, J.C. & Altschul, S.F. (2003). "The compositional adjustment of amino acid substitution matrices". *Proceedings of the National Academy of Sciences USA*, **100** (26), 15688-15693.

Zhang, C.T., Zhang, R. & Ou, H.Y. (2003). "The Z curve database: a graphic representation of genome sequences". *Bioinformatics*, **19** (5), 593-9.

Zhou, T. & Chiang, C.M. (2001). "The intronless and TATA-less human TAF(II)55 gene contains a functional initiator and a downstream promoter element". *Journal of Biological Chemistry*, **276** (27), 25503-11.